PROGRAMA DE DOCTORADO EN INGENIERÍAS

TESIS DOCTORAL:
EMBEBIMIENTO EN ESPACIOS DE HILBERT DE PROCESOS ALEATORIOS CON APLICACIONES EN PROCESAMIENTO DIGITAL DE SEÑALES

AUTOR:
EDGAR ALIRIO VALENCIA ANGULO

PEREIRA-2018

EMBEBIMIENTO EN ESPACIOS DE HILBERT DE PROCESOS ALEATORIOS CON APLICACIONES EN PROCESAMIENTO DIGITAL DE SEÑALES

EDGAR ALIRIO VALENCIA ANGULO

Tesis de grado presentado como requisito para optar por el título de Doctor en Ingeniería

Director: Ph.D Mauricio Alexánder Álvarez López

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
PROGRAMA DE DOCTORADO EN INGENIERÍAS
PEREIRA
2018

A mi madre Maria D. Angulo por su amor constante e incondicional; a mi esposa Claudia Andrea a mis hijos Lina Marcela, Ana Maria y Juan Camilo, deben saber que su apoyo y aliento valieron más de lo que puedo expresar en este papel; a mis hermanos Omar, Nilver Carmen y Ernesto por todos los consejos y motivaciones.

Declaración

Declaro que soy autor de la tesis con título Embebimiento en espacios de Hilbert de procesos aleatorios con aplicaciones en procesamiento digital de señales, que la tesis aquí contenida es mía, excepto donde explícitamente se indique lo contrario en el texto, y que esta tesis no se ha presentado para ningún otro título o calificación procesional en otra universidad.

Agradecimientos

Quiero agradecer a mi supervisor Mauricio A. Álvarez por su excelente guía, por los consejos y la motivación durante este proceso de aprendizaje. También quiero agradecer a la Universidad Tecnológica de Pereira y al Grupo de Investigación en Automática, más específicamente a mis compañeros Cristhian K. Valencia, Andrés F. López Lopera y Carlos D. Zuluaga por sus perspicaces comentarios y sugerencias.

Resumen

El método de embebimiento de distribuciones de probabilidad en un espacio de Hilbert con kernel reproductivo (RKHS) consiste en representar distribuciones de probabilidad como un elemento de un espacio de Hilbert generado por un kernel. Generalmente, este método ha sido usado en aplicaciones donde se supone que las observaciones son independientes e idénticamente distribuidas. Sin embargo, dentro de la literatura de los embebimiento de distribuciones de probabilidad en un RKHS, existen pocos trabajos donde se supone dependencia temporal de las observaciones. Motivado por este poderoso marco teórico del método de embebimiento de distribuciones en un RKHS, este trabajo de investigación desarrolla dos aplicaciones en Procesamiento Digital de Señales (DSP), donde se supondrá relación de dependencia entre las observaciones. En la primera aplicación, se introducen varias métricas entre distribuciones de probabilidad y entre modelos ocultos de Markov (HMMs), la discusión está limitada al kernel Gaussiano, al kernel de Laplaciano y al estimador de Parzen. Finalmente, se evalúa el rendimiento de las métricas en tareas de clasificación de series de tiempo, usando las métricas dentro del clasificador los K vecinos más cercanos. Los resultados muestran que nuestras métricas proporcionan una mejor precisión en clasificación de series de tiempo en comparación con la medida Kullback-Leibler (KL) y la métrica Euclidiana, en datos sintéticos y en datos reales. En la segunda aplicación, se propone una versión kernelizada de un modelo autoregresivo de orden p. Esta versión del modelo autorregresivo muestra un mayor rendimiento en predicción sobre el modelo lineal, en series de tiempo altamente complejas. Finalmente, la predicción se realiza un paso hacia adelante en diferentes series de tiempo, y nuestra versión del modelo autorregresivo se compara con otros métodos no lineales.

Abstrac

The method of embedding of probability distributions in a Reproducing Kernel Hilbert Space (RKHS) consists of representing probability distributions as an element of a Hilbert space generated by a kernel. Generally, this method has been used in applications where it is assumed that the observations are independent and identically distributed. However, within the literature of the probability distributions in a RKHS, there are few works where temporary dependence of the observations is assumed. Motivated by this powerful framework of the probability distributions in a RKHS, this research develops two applications in Digital Signal Processing (DSP), where dependence relationship between the observations will be assumed. In the first application, metrics between probability distributions and between stationary Hidden Markov Models (HMMs) using embedding of probability distributions in a RKHS are introduced, the discussion is limited to Gaussian kernel, Laplacian kernel and Parzen estimator. Finally, is evaluated the performance of the metrics in tasks of time series classification, using the metrics within the K-nearest neighbor classifier. The results show that our proposed metrics provides competitive in classification of time series accuracies when compared to the Kullback-Leibler (KL) and Euclidean metric in synthetic and real data. In the second application, a kernelized version of an autoregressive model of order p is proposed. This version of an autoregressive shows increased performance over the linear model in highly complex time series. Finally, One-step ahead forecasting of different time-series is made, and compared against other non-linear methods.

Tabla de contenido

Lista de figuras

Lista de tablas

Notación

Notación Matemática

Simbolos	Descripción
$\mathcal{X}$	Espacio de entrada
$\mathbb{R}$	Conjunto de números reales
$\mathcal{H}$	Espacio de Hilbert
$k(\cdot,\cdot)$	Función kernel
X	Variable aleatoria
$\mathcal{P}$	Espacio de todas las distribuciones de probabilidad
$\langle\cdot\ ,\cdot\rangle_{\mathcal{H}}$	Producto interno en $\mathcal{H}$
$\mathbb{P}$	Distribución de probabilidad
μ_X	Operador de media
$\mathbb{E}_X$	Esperanza de la variable aleatoria X
$\mathcal{F}$	Sigma-álgebra
$\{x^l\}_{l=1}^m$	Conjunto de N_X observaciones *i.i.d.* de la variable aleatoria X
$\widehat{\mu}_X$	Estimador de μ_X
$f \otimes g$	Producto tensor entre las funciones f y g
$\mathcal{H}_1 \otimes \mathcal{H}_2$	Espacio producto tensor entre los espacios de Hilbert $\mathcal{H}_1$ y $\mathcal{H}_2$
$\mathcal{P}_{XY}$	Conjunto de distribuciones de probabilidad conjuntas
$\mathcal{C}_{XY}$	Operador de covarianza cruzada entre las variables aleatorias X y Y
$\{(x^i,y^i)\}_{i=1}^m$	Conjunto de N_{XY} observaciones *i.i.d.* del vector aleatorio (X,Y)
$\widehat{\mathbf{C}}_{XY}$	Estimador de $\mathcal{C}_{XY}$
$\boldsymbol{\Upsilon}$ y $\boldsymbol{\Phi}$	Matrices características
$\mathbf{K}$, $\mathbf{L}$ y $\mathbf{H}$	Matrices kernel
$\mathrm{tr}(\cdot)$	Traza de una matriz
$\gamma_k(\mathbb{P},\mathbb{Q})$	Métrica entre las distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$
$d_\gamma(\mathbb{P},\mathbb{Q})_{Laplaciana}$	Métrica entre las distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$ usando el kernel Laplaciano
$d_\gamma(\mathbb{P},\mathbb{Q})_{Parzen}$	Métrica entre las distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$ usando el estimador de Parzen
$\delta(\cdot)$	Función delta Dirac
$\mathcal{M}_{m\times m}$	Conjunto de todas las matrices de orden $m \times m$
$\mathrm{diag}(\cdot)$	Diagonal de una matriz
$\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$ y $\mathbf{b}$	Vectores
$\mathbf{I}$	Matrix identidad
$\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma})$	Distribución normal con vector media $\boldsymbol{\mu}$, y matriz de covarianza $\boldsymbol{\Sigma}$
$\mathrm{erf}(\cdot)$	Función Error

Tabla 1: Notación y simbolos utilizados en la tesis

Abreviaciones	**Descripción**
PDS	Procesamiento digital de señales (digital signal processing)
RKHS	Espacio de Hilbert con kernel reproductivo (reproducing kernel Hilbert space)
TP-RKHS	Producto Tensor del Espacio de Hilbert con kernel reproductivo (Tensor product reproducing kernel)
i.i.d.	Independiente e idénticamente distribuidas
AR	Modelo autorregresivo (Autoregressive model)
LAR	Modelo AR Lineal
MAK	Modelo Autorregresivo Kernel
MEK	Modelo Embebimiento Kernel
GP	Procesos Gaussianos (Gaussian processes)
NN	Redes Neuronales (Neural networks)
MMD	Máxima discrepancia en media (maximun mean discrepancy)
HMM	Modelo oculto de Markov (Hidden Markov Model)
KNN	K-vecino más cercano (K-Nearest Neighbors)
EUC	Métrica Euclideana
PKE	Métrica basada en el estimador de Parzen y en método de embebimiento de distribuciones de probabilidad en un RKHS
KL	Kullback-Leibler
TSDL	Librería de datos de series de tiempo (Time Series Data Library)
CV	Coeficiente de variación
HSD	Medida de distancia para modelos ocultos de Markov estacionarios (HMMs stationary distance measure)
MSE	Error cuadrático medio (mean squared error)
KEG	Métrica basada en Embebimiento con kernel Gaussiano
KEL	Métrica basada en Embebimiento con kernel Laplaciano

Tabla 2: Abreviaciones utilizadas en la tesis

Capítulo 1

Introducción

Los métodos kernel son una familia de algoritmos de aprendizaje de máquina ampliamente utilizados en procesamiento digital de señales (DSP) [17]. Su popularidad se puede atribuir a su sólida base matemática dentro de los espacios de Hilbert generados por kernel y porque han demostrado tener buen desempeño en la solución de problemas no lineales [39, 54]. Debido a estas propiedades, los métodos kernel representan una alternativa a los métodos tradicionales no lineales como las redes neuronales artificiales, las máquinas de vectores de soporte lineales y los filtros de polinomios. Sin embargo, en algunas áreas como: pruebas estadísticas para distribuciones de probabilidad, estimación de distancia entre distribuciones de probabilidad, medidas de dependencia entre variables aleatorias y medidas de dependencia entre procesos aleatorios, estos métodos no se han podido consolidar [11].

Dentro de la literatura de los métodos kernel, recientemente el método embebimiento de distribuciones de probabilidad en espacios de Hilbert con kernel reproductivo [1] (RKHS) ha sido usado para mapear distribuciones en un espacio de Hilbert. Este método consiste en la representación de distribuciones de probabilidad como puntos en un RKHS a través de un operador inyectivo. Algunas de las razones por las que ha sido usado el método de embebimiento son por sus diversas propiedades, como por ejemplo: permiten modelar datos sin necesidad de hacer suposiciones sobre el tipo de distribuciones de probabilidad; pueden ser aplicados en cualquier dominio en el que se puede definir un kernel y permiten aplicar conceptos como tensores para realizar tareas de aprendizaje dando lugar a métodos de extracción de características y a métodos de estimación de parámetros [45].

En la solución de problemas de DSP usando el método embebimiento de distribuciones de probabilidad en un RKHS, es necesario tener un estimador consistente de una medida de similaridad o una medida de distancia entre distribuciones de probabilidad para obtener un buen rendimiento en estos problemas. La consistencia del estimador es garantizada si y sólo si las observaciones son independientes e idénticamente distribuidas [41, 14]. Sin embargo, en aplicaciones de HMMs y AR, las observaciones no satisfacen

[1] Las palabras kernel reproductivo y embebimiento, también se traducen como nucleo reproduciente y encaje (o incrustación) respectivamente, pero nosotros escogemos kernel reproductivo y embebimieto porque son palabras más comunes en la literatura de los procesos aleatorios aplicados en DSP.

la suposición de independencia e idénticamente distribuidas ($i.i.d$).

Motivado por este poderoso marco, en este trabajo de investigación desarrollamos diferentes métodos y algoritmos que permitirán realizar la predicción en un proceso autorregresivo usando el método embebimiento de distribuciones de probabilidad en un RKHS. Consideramos que hacer predicción en un proceso autorregresivo usando el método embebimiento de distribuciones de probabilidad conjuntas en un RKHS es una investigación novedosa. Existe un número reducido de métodos basados en kernel que realizan pronóstico en modelos autorregresivos con buen rendimiento, incluyendo el enfoque de modelado de kernel autorregresivo [24], modelos autorregresivos basados en kernel usando las ecuaciones de Yule-Walker [23] y el problema de la pre-imagen [18], predicción de series cronológicas con kernel [37] y kernel con filtros de Kalman [35]. Dado que el método embebimiento de distribuciones de probabilidad conjuntas mide de manera eficiente la relación de dependencia entre variables aleatorias y es una generalización de los métodos tradicionales basados en kernel, se considera que es posible mejorar la estimación y predicción de un modelo autorregresivo. En este sentido, nuestra investigación es importante porque utiliza la reciente metodología de embebimiento de distribuciones de probabilidad conjuntas en un RKHS [45], las ecuaciones de Yule-Walker y el problema de la pre-imagen [18] con el propósito de mejorar el rendimiento en la estimación y predicción de un AR.

Otra de las aplicaciones que vamos a realizar en este trababjo de investigación usando la teoría de embebimiento de distribuciones de probabilidad en un RKHS, es calcular medidas de distancia entre dos procesos aleatorios con estructura Markoviana. En problemas de clasificación de procesos aleatorios se han propuesto muchas medidas de similitud como por ejemplo, la distancia Euclidiana entre matrices de datos de distribuciones de probabilida [21], la distancia de Bhattacharyya [2], la divergencia de Kullback-Leibler KL y sus modificaciones [57], medidas basadas en probabilidades de emisión y medidas basadas en el error de la probabilidad de Bayes [50]. Sin embargo, algunas medidas de distancia como KL y Bhattacharyya no son verdaderas métricas, por lo tanto estas medidas pueden producir bajo rendimiento en problemas de clasificación [11, 58]. El trabajo presentado en [56] propone una verdadera medida de distancia basada en la distribución de probabilidad estacionaria de un HMM. Una desventaja de esta medida de distancia es que se define en espacios de una dimensión y su cálculo en espacios de alta dimensión puede ser complicado.

1.1 Estado del arte

El método embebimiento de distribuciones de probabilidad en un RKHS, hace parte de los métodos basados en kernel. Este método ha sido desarrollado principalmente en el año 2007 en [44]. La idea básica de este método es que mapea distribuciones de probabilidad en un espacio de características de alta dimensión. En los últimos tiempos el método embebimiento de distribuciones de probabilidad en un RKHS, ha sido utilizado con gran éxito como alternativa a los tradicionales modelos probabilísticos paramétricos para reducción de dimensionalidad [12], regla de Bayes con kernel [13] y estimación de distancia entre distribuciones de probabilidad [14].

Algunas de las razones por las que ha sido usado este método son por sus diversas propiedades: permite modelar datos sin necesidad de hacer suposiciones sobre el tipo de distribuciones de probabilidad; pueden ser aplicados en cualquier dominio en el que se puede definir un kernel y permite aplicar conceptos como tensores para realizar tareas de aprendizaje dando lugar a métodos de extracción de características y a métodos de estimación de parámetros [47].

Cuando se ha utilizado el método embebimiento de distribuciones de probabilidad en un RKHS en la solución de problemas de DSP, no se han tenido grandes dificultades en el cálculo de la estimación de parámetros de los procesos aleatorios, si se compara con los métodos tradicionales de estimación de estos procesos [46], lo contrario sucede con la predicción de procesos aleatorios usando los embebimientos en un RKHS. Esto es debido a que la predicción se hace en el espacio de entrada y la estimación de los parámetros se hace en el embebimiento RKHS. Sin embargo, si el kernel es Gaussiano y es definido en un conjunto compacto, el embebimiento puede verse como un estimador de densidad no paramétrico que permite encontrar las predicciones [46].

El método embebimiento de distribuciones en un RKHS, también ha sido aplicado a pruebas de independencia [14], pruebas de homogeneidad [49], pruebas de independencia condicional [14] y pruebas de dos muestras [6]. En todas estas pruebas, se supone que las observaciones son *i.i.d.* Todas las aplicaciones de los embebimientos de distribuciones de probabilidad en un RKHS se basan en el operador de covarianza cruzada, este operador depende básicamente de un kernel característico [47] y su estimador (la covarianza empírica) tiene la propiedad de ser un estimador consistente para muestras de gran tamaño. Sin embargo, cuando la muestra es pequeña el estimador del operador covarianza cruzada no es consistente [36]. Para solucionar este inconveniente, en [29] se propone dos estimadores de contracción y en [36] se estudia el fenómeno de la contracción del operador de covarianza cruzada en RKHS, se desarrolla una tercera familia de estimadores de contracción y se realiza un estudio de cómo la contracción mejora la estimación del operador de covarianza cruzada.

En la mayoría de las aplicaciones de los embebimientos de distribuciones en un RKHS, se supone que las observaciones son independientes e idénticamente distribuidas. No obstante, las observaciones del mundo real generalmente no cumplen el supuesto de independencia e idénticamente distribuidas, como por ejemplo en: señales de audio, documentos de texto, series de tiempo, y las muestras obtenidas de los métodos de cadenas de Markov Monte Carlo, todos estos ejemplos muestran significativos patrones de dependencia temporales [9]. Además, suponer dependencia temporal en las observaciones usando el método embebimiento de distribuciones de probablidad en un RKHS, permite obtener un comportamiento asintótico de las pruebas estadísticas [9].

1.2 Planteamiento del problema

Normalmente, el método embebimiento de distribuciones de probabilidad en un RKHS ha sido usado en aplicaciones donde se supone que las observaciones son *i.i.d.*. Sin embargo, dentro de la literatura de los embebimientos de distribuciones de probabilidad en un RKHS, existen pocos trabajos donde se supone dependencia temporal de las

observaciones [9, 46, 45]. En [46] los autores proponen un embebimiento en un RKHS de un HMMs, que extiende los tradicionales HMMs a observaciones distribuidas de forma no Gaussiana y a observaciones con dependencia temporal. En [45] los autores proponen un marco general para el método embebimiento de distribuciones condicionales en un RKHS, y muestran como este método es útil para modelar y realizar inferencia no paramétrica sobre sistemas dinámicos. Finalmente en [9], los autores proponen una medida de distancia entre procesos aleatorios usando la medida de distancia MMD propuesta en [14] y el método de remuestreo de bootstrap. Dicho de otra manera, el método embebimiento de distribuciones de probabilidad en un RKHS, ha permitido resolver algunos problemas de DSP donde se mide la relación de dependencia entre variables aleatorias y entre procesos aleatorios en espacios de alta dimensión de forma eficiente. Por lo tanto, la pregunta que surge en este trabajo de investigación es la siguiente: ¿es posible usar el método de embebimiento de distribuciones de probabilidad en un RKHS en la solución de otros problemas de DSP que midan la relación de dependencia entre procesos aleatorios?. En este trabajo de investigación, se explorarán dos aplicaciones en DSP, usando el método embebimiento de distribuciones de probabilidad en un RKHS, donde se supondrá relación de dependencia entre las observaciones, estas aplicaciones son: la estimación y predicción de un modelo autorregresivo y la construcción analítica de nuevas medidas de distancia entre HMMs.

1.3 Contribución de la tesis

En este trabajo de investigación se desarrollan diferentes métodos y algoritmos basados en el método embebimiento de procesos aleatorios discretos en un RKHS con aplicaciones en predicción y clasificación de series de tiempo. Especialmente, se desarrolla una extensión no lineal de un proceso autorregresivo aplicado a la predicción de series de tiempo. También se desarrollan de forma analítica métricas entre distribuciones de probablidad y entre HMMs, con aplicaciones en clasificación de series de tiempo.

Las aplicaciones que se desarrollan en este trabajo, usando el método embebimiento de distribuciones de probabilidad en un RKHS tienen en común la dependencia temporal de las observaciones. Estas aplicaciones están relacionas entre sí (ver Figura 1.1), por ejemplo una medida de dependencia entre variables aleatorias es un caso particular de una métrica entre distribuciones de probabilidad. Por otro lado, a partir de la construcción de una métrica entre distribuciones de probabilidad, podemos desarrollar una métrica entre HMMs estacionarios. Sin embargo, una pregunta que surge de esta investigación es la siguiente: ¿es posible construir métricas entre procesos aleatorios distintos a los HMMs usando el método embebimiento de distribuciones de probabilidad en un RKHS?. Creemos que sí es posible, siguiendo procedimientos basados en el que presentamos es esta tesis. La construcción de métricas entre otros procesos aleatorios distintos al HMM, serían trabajos futuros a desarrollar.

1.4 Publicaciones

Las contribuciones de la tesis han sido presentadas en las siguientes publicaciones en revistas indexadas categoría A.

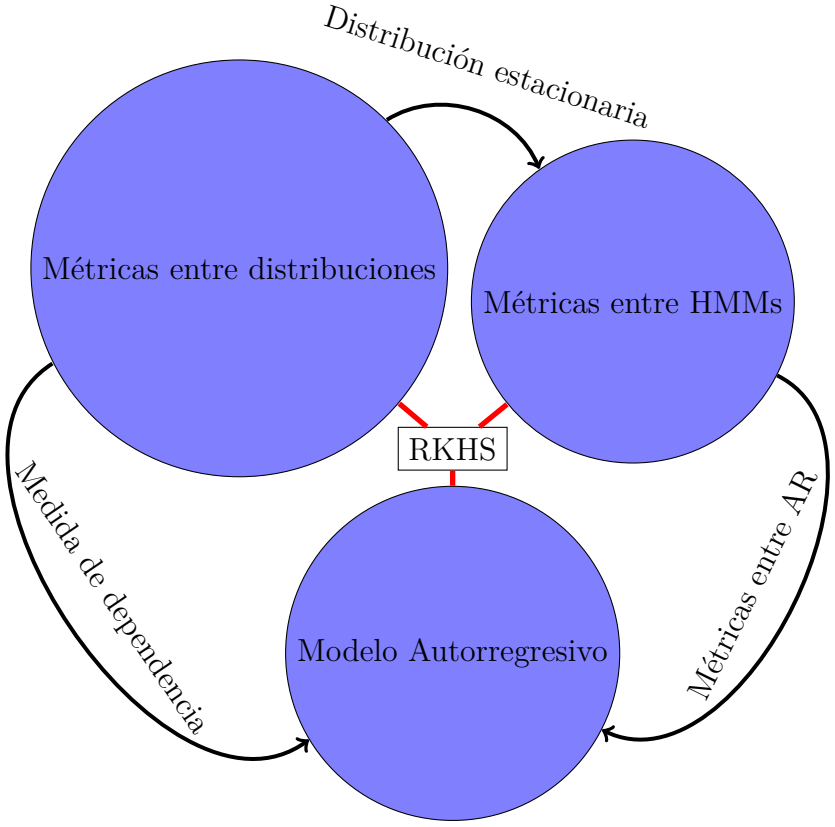


Figura 1.1: Relación entre las aplicaciones en DSP presentadas en la tesis

(i) Carlos D. Zuluaga, Edgar A. Valencia, Mauricio A. Álvarez, Álvaro Á. Orozco: A Parzen-Based Distance Between Probability Measures as an Alternative of Summary Statistics in Approximate Bayesian Computation. Lecture Notes in Computer Science. International Conference on Image Analysis and Processing (ICIAP), 50-61, 2015, Springer.

(ii) E. A. Valencia, Cristhian K. Valencia, Andrés F. Lópera and M. A. Álvarez: Distance measures for hidden Markov models based on Hilbert space embeddings for time series classification, Pattern Recognition Letters (el artículo está en revisión).

(iii) E. A. Valencia and M. A. Álvarez (2017): Short-term time series prediction using Hilbert space embeddings of autoregressive processes, Neurocomputing, volume 266, pages 595-605, año 2017.

1.5 Estructura de la tesis

Este trabajo está organizado de la siguiente manera: el Capítulo 2 describe el método embebimiento de distribuciones de probabilidad en un RKHS, se presenta la definición de

los RKHS y sus propiedades más importantes. El Capítulo 3 describe las métricas entre las distribuciones de probabilidad y entre los modelos ocultos de Markov basados en los embebimientos de distribuciones de probabilidad en un RKHS, para la clasificación de series de tiempo. En este capítulo presentamos resultados experimentales utilizando las métricas basadas en los RKHS. El Capítulo 4 describe la predicción de series de tiempo a corto plazo utilizando embebimientos en espacios de Hilbert de procesos autorregresivos, en este capítulo se hace predicción un paso hacia adelante para cuatro conjuntos de datos. En el Capítulo 5, se presentan las conclusiones y trabajos futuros, derivados de esta investigación. Finalmente, se presentan las referencias y los apéndices, donde estan las pruebas de los teoremas más importantes de esta investigación.

Capítulo 2

Embebimiento de distribuciones de probabilidad en un RKHS

Un método alternativo y reciente de estimación no paramétrica, consiste en embeber un conjunto de distribuciones de probabilidad o distribuciones de probabilidad condicionales en un RKHS. Este método ha surgido como una poderosa herramienta para la elaboración de modelos probabilísticos, inferencia estadística y aprendizaje automático [29]. Así mismo, el método embebimiento de distribuciones de probabilidad en un RKHS comprende una clase de métodos no paramétricos en los cuales una distribución de probabilidad es caracterizada como un elemento de un RKHS [44]. La idea fundamental de este método consiste en representar distribuciones de probabilidad como un elemento de un espacio de Hilbert generado por un kernel, tal que las operaciones, las comparaciones, y las manipulaciones de estas distribuciones de probabilidad se llevan a cabo en este espacio de Hilbert [47]. Una condición que debe cumplir el kernel que genera el espacio de Hilbert es que sea característico, es decir el operador que permite el embebimiento del conjunto de distribuciones de probabilidad al espacio de Hilbert (espacio de característica), es un operador inyectivo.

Para realizar el embebimiento del conjunto de distribuciones de probabilidad conjuntas al conjunto RKHS, se define el operador de covarianza el cual está expresado en términos de un producto exterior (ver [1]) y cuyo estimador se calcula a partir de un conjunto de observaciones $i.i.d$ (ver [46]).

En este capítulo repasamos brevemente las definiciones de un espacio de Hilbert con kernel reproductivo, embebimiento de distribuciones de probabilidad en un RKHS, operadores de covarianza y sus propiedades y medidas de distancia entre distribuciones de probabilidad en un RKHS.

En este trabajo, se usan letras mayúsculas para referirse a variables aleatorias (por ejemplo, X, Y), letras minúsculas para referirse a los valores particulares que estas variables aleatorias puedan tomar (por ejemplo, x, y), letras minúsculas en negrilla para referirse a vectores (por ejemplo $\mathbf{x}, \mathbf{y}$) y letras mayúsculas en negrilla para referirse a matrices (por ejemplo $\mathbf{A}, \mathbf{B}$).

2.1 Definición y propiedades de un RKHS

En esta subsección presentamos la definición y algunas propiedades básicas de un RKHS.

Definición 2.1.1. *Dado el espacio topológico $\mathbb{R}$, y una función $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, llamada kernel reproductivo donde $\mathcal{X}$ es cualquier conjunto. Un espacio de Hilbert $\mathcal{H}$ con kernel reproductivo $k(x,y)$, para $x,y \in \mathcal{X}$, es un espacio de funciones $g : \mathcal{X} \to \mathbb{R}$ que satisface las siguientes propiedades:*

1. *Para todo $x \in \mathcal{X}$, $k_x : \mathcal{X} \to \mathbb{R}$ donde $k_x \in \mathcal{H}$. Es decir, $k_x : y \to k(x,y)$, $k_x(y) = k(x,y)$.*

2. *Propiedad de reproducción $g(x) = \langle g, k_x \rangle_{\mathcal{H}}$.*

A continuación presentamos una propiedad de un RKHS que es una consecuencia de la definición 2.1.1.

Proposición 2.1.2. *Sea $\mathcal{H}$ un RKHS, y $\phi : \mathcal{X} \to \mathcal{H}$, entonces*

$$k(x,y) = \langle k_x, k_y \rangle_{\mathcal{H}} = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} \quad \text{donde } k_x = \phi(x), \tag{2.1}$$

es un kernel reproductivo. Además, este kernel es único.

Prueba Verifiquemos que k es simétrica y definida positiva.

Mostremos que k es simétrica.

$$k(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \langle \phi(y), \phi(x) \rangle_{\mathcal{H}} = k(y,x).$$

Sea $x_1, x_2, \cdots x_n \in \mathcal{X}$ y $\alpha_1, \alpha_2, \cdots \alpha_n \in \mathbb{R}$, mostremos que k es definida positiva.

$$\begin{aligned}\sum_{j=1}^{n}\sum_{i=1}^{n} \alpha_i \alpha_j k(x_i, x_j) &= \sum_{j=1}^{n}\sum_{i=1}^{n} \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) = \sum_{j=1}^{n}\sum_{i=1}^{n} \langle \alpha_i \phi(x_i), \alpha_j \phi(x_j) \rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{n} \alpha_i \phi(x_i), \sum_{j=1}^{n} \alpha_j \phi(x_j) \right\rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} \alpha_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0.\end{aligned}$$

Finalmente, mostremos la unicidad del kernel reproductivo. Supongamos que k y k' son dos kernels reproductivo de $\mathcal{H}$ y $x,y \in \mathcal{X}$, necesitamos demostrar que $k(x,y) = k'(x,y)$. Por definición de kernel reproductivo $k_x, k'_x \in \mathcal{H}$, luego

$$\begin{aligned}\|k_x - k'_x\|_{\mathcal{H}}^2 &= \left\langle k_x - k'_x, k_x - k'_x \right\rangle_{\mathcal{H}} = \left\langle k_x - k'_x, k_x \right\rangle_{\mathcal{H}} - \left\langle k_x - k'_x, k'_x \right\rangle_{\mathcal{H}} \\ &= (k_x - k'_x)(x) - (k_x - k'_x)(x) = k_x(x) - k'_x(x) - k_x(x) + k'_x(x) = 0.\end{aligned}$$

Por lo tanto $k_x = k'_x$, en particular $k_x(y) = k'_x(y)$, es decir $k(x,y) = k'(x,y)$.□

Una definición alternativa para una función kernel, que suele ser usada cuando se diseñan algoritmos, es dado por la expresión 2.1.

2.2 Embebimiento de distribuciones de probabilidad

En [44] se introdujo un método para encajar distribuciones de probabilidad en un RKHS.

Definición 2.2.1. *Sea $\mathcal{P}$ el espacio de todas las distribuciones de probabilidad sobre un espacio de medida $(\mathcal{X},\mathcal{F})$ donde $\mathcal{F}$ es una sigma-álgebra de $\mathcal{X}$, y X una variable aleatoria con distribución de probabilidad $\mathbb{P}\in\mathcal{P}$. En [44], los autores definen el mapeo de la distribución de probabilidad $\mathbb{P}\in\mathcal{P}$ a un RKHS $\mathcal{H}$ usando el operador media μ_X definido como*

$$\mu_X = \mathbb{E}_X[k_X] = \mathbb{E}_X[\phi(X)],$$

donde $\mathbb{E}_X$ es la esperanza de la variable aleatoria X.

Note que el operador de media del embebimiento μ_X satisface $\langle \mu_X, \varphi(\cdot)\rangle_{\mathcal{H}} = \mathbb{E}_X[\varphi(X)]$ y el kernel $k(x,x')$ usado para el embebimiento es característico, es decir cumple la siguiente definición:

Definición 2.2.2. *Sea $\mathcal{P}$ el espacio de todas las distribuciones de probabilidad sobre un espacio de medida $(\mathcal{X},\mathcal{F})$ donde $\mathcal{F}$ es una sigma-álgebra de $\mathcal{X}$. Un kernel característico es un kernel reproductivo tal que si*

$$\mu_X(\mathbb{P}) = \mu_Y(\mathbb{Q}) \quad \textit{entonces} \quad \mathbb{P}=\mathbb{Q}, \quad \textit{para} \quad \mathbb{P},\ \mathbb{Q}\in\mathcal{P}.$$

Es decir, μ_X es un operador inyectivo.

Ejemplo 2.2.3. *Algunos ejemplos de kernels característico sobre el conjunto $\mathcal{X}=\mathbb{R}^d$ son:*

(a) Kernel Gaussiano $k(x,y) = \exp\left(-\sigma\|x-y\|_2^2\right)$ donde $\sigma>0$.

(b) Kernel Laplaciano $k(x,y) = \exp\left(-\sigma\|x-y\|_1\right)$ donde $\sigma>0$.

(c) Kernel inverso multiplicativo $k(x,y) = \left(\sigma^2+\|x-y\|_2^2\right)^{-c}$ donde $c,\ \sigma>0$.

(d) Kernel B_{2n+1}-splines $k(x,y) = (1-|x-y|)\,1_{[-1,1]}(x-y)$ donde $x,\ y\in\mathbb{R}$.

(e) Kernel Dirichlet $k(x,y) = \frac{\sin^2\frac{((\ell+1)(x-y))}{2}}{\sin^2\frac{(x-y)}{2}}$ donde $\ell\in\mathbb{N}$.

En este trabajo de investigación, usaremos el kernel Gaussiano y el kernel Laplaciano, dado que nos permiten calcular de forma cerrada métricas entre distribuciones de probabilidad y entre HMMs. Otra razón para usar el kernel Gaussiano es por que tiene la propiedad de ser un kernel diferenciable, el cual nos permitirá hacer predicción en un modelo autorregresivo encajado en un TP-RKHS.

La Figura 2.1 muestra los dos tipos de kernels característicos utilizados en este trabajo: kernel Gaussiano y kernel Laplaciano. Podemos observar que las colas del kernel Gaussiano son más pesadas que la del kernel Laplaciano, dando mucha menos probabilidad a los valores lejos de la media. También observamos que, a diferencia del kernel Laplanciano, el kernel Gaussiano es más suave. Esta propiedad puede explotarse para describir distribuciones de probabilidad diferenciables.

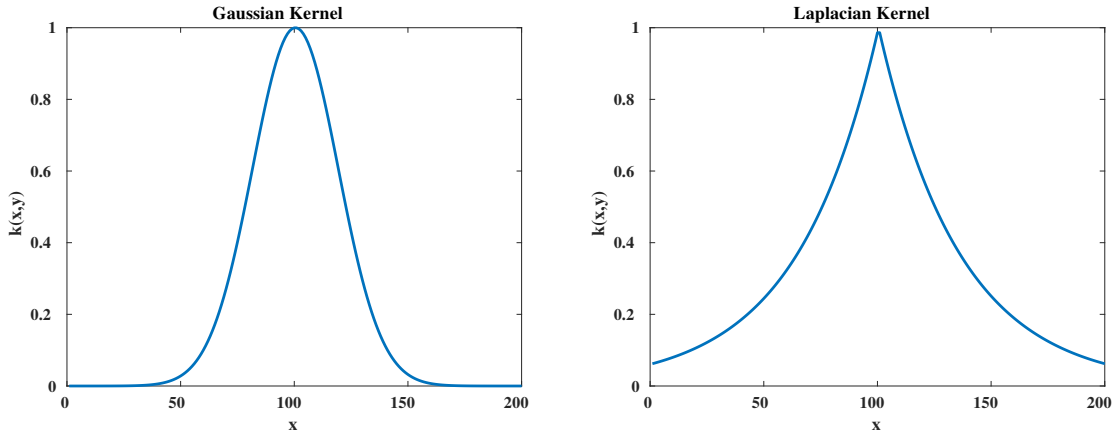


Figura 2.1: Gráfica del kernel Gaussiano y el kernel Laplaciano.

Definición 2.2.4. *Dado un conjunto de observaciones i.i.d.* $\{x_i\}_{i=1}^{N_x}$ *de tamaño* N_x *de la variable aleatoria* X*, un estimador para* μ_X *está dado como:*

$$\widehat{\mu}_X = \frac{1}{N_x}\sum_{i=1}^{N_x} k_{x_i}.$$

El estimador $\widehat{\mu}_X$, tiene las siguientes propiedades:

(a) $\langle \widehat{\mu}_X, \phi(\cdot)\rangle_{\mathcal{H}} = \frac{1}{N_x}\sum_{i=1}^{N_x}\phi(x_i)$ donde $\phi(x_i) \in \mathcal{H}$.

(b) El estimador $\widehat{\mu}_X$ converge a μ_X, en la norma definida en $\mathcal{H}$, con rapidez de convergencia de $O(N_x^{-1/2})$ es decir, exite un $\epsilon > 0$ tal que $\|\mu_X - \widehat{\mu}_X\| \leq \epsilon N_x^{-1/2}$ para N_x muy grande.

Las pruebas se pueden ver en [44].

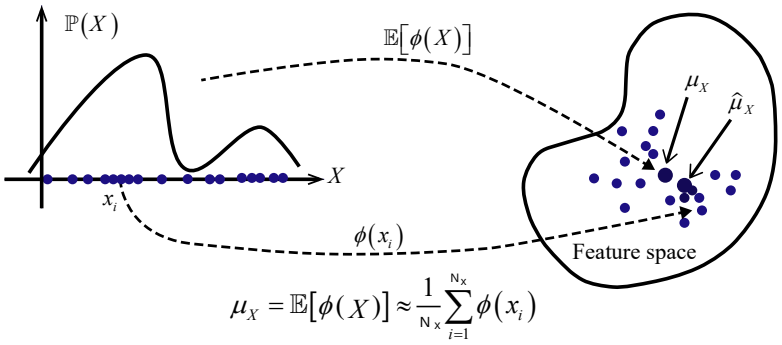


Figura 2.2: El operador de embebimiento de distribuciones de probabilidad y su estimador usando una muestra finita, donde N_x es el tamaño de la muestra. Imagen tomada de [47] y editada por el autor.

La Figura 2.2 muestra el embebimiento de distribuciones de probabilidad en un espacio de característica o espacio de Hilbert. La Figura 2.2 también muestra la relación entre el operador de embebimiento de distribuciones de probabilidad y su estimador.

2.3 Operador de covarianza cruzada

Los operadores de covarianza cruzada se introdujeron en [1] como una herramienta en la solución de problemas de medidas de probabilidades definidas sobre el producto tensorial de dos espacios de Hilbert separables.

Definición 2.3.1. *Dado $f,h \in \mathcal{H}_1$, y $g \in \mathcal{H}_2$, se define el producto tensorial $f \otimes g$ como un operador que mapea cada elemento h de $\mathcal{H}_1$ a un elemento de $\mathcal{H}_2$,*

$$(f \otimes g)h \longrightarrow \langle h, f\rangle_{\mathcal{H}_1} g. \tag{2.2}$$

Definición 2.3.2. *Sean $\mathcal{H}_1$ y $\mathcal{H}_2$ dos RKHS con kernels $k(\cdot,\cdot)$ y $\ell(\cdot,\cdot)$, y mapeos de características ϕ y φ respectivamente, el operador de covarianza cruzada está definido como en [1]*

$$\mathcal{C}_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \varphi(Y)]. \tag{2.3}$$

El operador de covarianza cruzada $\mathcal{C}_{XY}$ puede ser visto como a elemento de un producto tensor del espacio de Hilbert con kernel reproductivo (TP-RKHS), $\mathcal{H}_1 \otimes \mathcal{H}_2$ [14].

Observación 2.3.3. *Dadas dos funciones $f \in \mathcal{H}_1$ y $g \in \mathcal{H}_2$, entonces*

$$\begin{aligned}
\langle f, \mathcal{C}_{XY} g\rangle_{\mathcal{H}_1} &= \langle f \otimes g, \mathcal{C}_{XY}\rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2} \\
&= \mathbb{E}_{XY}\left[\langle f \otimes g, \phi(X) \otimes \varphi(Y)\rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2}\right] \\
&= \mathbb{E}_{XY}\left[\langle f,\ \phi(X)\rangle_{\mathcal{H}_1} \langle g,\ \varphi(Y)\rangle_{\mathcal{H}_2}\right] \\
&= \mathbb{E}_{XY}\left[f(X)g(Y)\right],
\end{aligned}$$

donde $\phi(x) = k(x,\cdot)$, $\varphi(y) = l(y,\cdot)$, y $\mathbb{E}_{XY}[f(x)g(y)]$ es la matriz de covarianza. Para más detalles ver [14].

El operador $\mathcal{C}_{XY}$ permite el embebimiento de $\mathcal{P}_{XY}$, el conjunto de distribuciones conjuntas $\mathbb{P}(X,Y)$ sobre el vector aleatorio $(X,Y) \in \mathcal{X} \times \mathcal{Y}$, en el TP-RKHS $\mathcal{H}_1 \otimes \mathcal{H}_2$.

Proposición 2.3.4. *Si $\mathcal{C}_{XY} = \mathbb{E}_{XY}[\varphi(X) \otimes \phi(Y)]$, $\varphi(X) = k_X$ y $\phi(Y) = l_Y$ donde k y l son funciones kernels, entonces*

$$\|\mathcal{C}_{XY}\|^2_{\mathcal{H}_1 \otimes \mathcal{H}_2} = \mathbb{E}_{XY}\mathbb{E}_{X'Y'}[k(X,X')l(Y,Y')]. \tag{2.4}$$

Prueba

$$\begin{aligned}
\|\mathcal{C}_{XY}\|^2_{\mathcal{H}_1 \otimes \mathcal{H}_2} &= \langle \mathcal{C}_{XY}, \mathcal{C}_{XY}\rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2} = \mathbb{E}_{XY}\left[\langle \varphi(X) \otimes \phi(Y), \mathcal{C}_{XY}\rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2}\right] \\
&= \mathbb{E}_{XY} E_{X'Y'}\left[\langle \varphi(X) \otimes \phi(Y), \varphi(X') \otimes \phi(Y')\rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2}\right] \\
&= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\left[\langle \varphi(X), \varphi(X')\rangle_{\mathcal{H}_1} \langle \phi(Y), \phi(Y')\rangle_{\mathcal{H}_2}\right] \\
&= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}[k(X,X')l(Y,Y')]. \ \square
\end{aligned} \tag{2.5}$$

Definición 2.3.5. *Considere un conjunto de pares de observaciones i.i.d $\{(x_i,y_i)\}_{i=1}^{N_{xy}}$, de un vector aleatorio (X,Y), un estimador de covarianza cruzada $\widehat{\mathbf{C}}_{XY}$ para $\mathcal{C}_{XY}$ se*

define como

$$\widehat{\mathbf{C}}_{XY} = \frac{1}{N_{xy}} \sum_{i=1}^{N_{xy}} \phi(x_i) \otimes \varphi(y_i) = \frac{1}{N_{xy}} \mathbf{\Phi} \mathbf{\Upsilon}^\top, \tag{2.6}$$

donde $\mathbf{\Phi} = (\phi(x_1), \phi(x_2), \ldots, \phi(x_{N_{xy}}))$, *y* $\mathbf{\Upsilon} = (\varphi(y_1), \varphi(y_2), \ldots, \varphi(y_{N_{xy}}))$ *son matrices de características [4]*

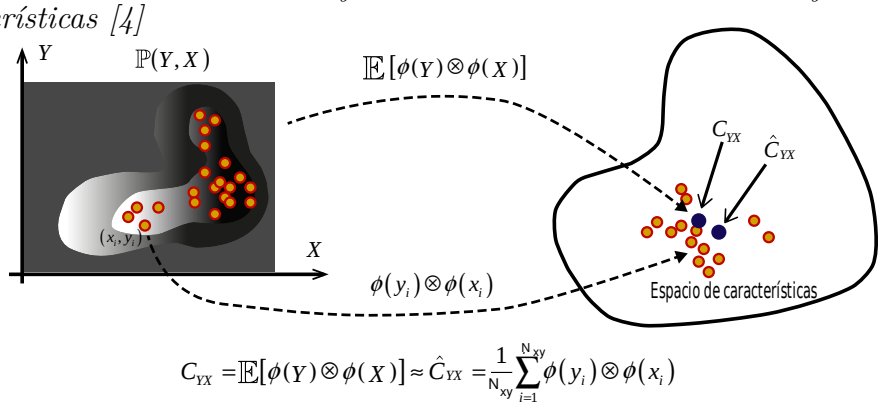


Figura 2.3: Operador de embebimiento de distribuciones de probabilidad conjuntas y su estimador usando una muestra finita. Imagen tomada de [47] y editada por el autor.

La Figura 2.3 muestra el embebimiento de distribuciones de probabilidad conjunta en el espacio de Hilbert TP-RKHS. La Figura 2.3 también muestra la relación entre el operador de embebimiento de distribuciones de probabilidad conjunta y su estimador.

Proposición 2.3.6. *Si* $\widehat{\mathbf{C}}_{XY} = \frac{1}{N_{xy}} \mathbf{\Upsilon} \mathbf{\Phi}^\top$, *entonces*

$$\|\widehat{\mathbf{C}}_{XY}\|^2 = \frac{1}{N_{xy}^2} \operatorname{tr}(\mathbf{KL}), \tag{2.7}$$

donde $k_{ij} = k(x_i, x_j)$, $l_{ij} = l(y_i, y_j)$, $\mathbf{K} = (k_{ij})$ *y* $\mathbf{L} = (l_{ij})$ *respectivamente.*

Prueba

$$\begin{aligned} \|\widehat{\mathbf{C}}_{XY}\|^2 &= \left\langle \frac{1}{N_{xy}} \sum_{i=1}^{N_{xy}} \varphi(x_i) \otimes \phi(y_i), \frac{1}{N_{xy}} \sum_{i=1}^{N_{xy}} \varphi(x_i) \otimes \phi(y_i) \right\rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2} \\ &= \frac{1}{N_{xy}^2} \sum_{j=1}^{N_{xy}} \sum_{i=1}^{N_{xy}} k(x_i, x_j) l(y_i, y_j) = \frac{1}{N_{xy}^2} \operatorname{tr}(\mathbf{KL}). \ \square \end{aligned} \tag{2.8}$$

2.4 Medidas de distancia entre distribuciones de probabilidad en un RKHS

El concepto de medida de distancia o métrica entre distribuciones de probabilidad y entre procesos aleatorios, es fundamental y ha encontrado muchas aplicaciones en la teoría de la probabilidad y estadística y la teoría de la información [19, 27]. En esta sección se estudia la medida de similaridad entre distribuciones de probabilidad definida

por Muller en [30].

Definición 2.4.1. *Sea $\mathcal{P}$ el conjunto de todas las distribuciones de probabilidad de Borel sobre $(\mathcal{X},\mathcal{F})$, la medida de similaridad entre $\mathbb{P}\in\mathcal{P}$ y $\mathbb{Q}\in\mathcal{P}$ se define como*

$$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})=\sup_{f\in\mathcal{F}}\left|\int_{\mathcal{X}}fd\mathbb{P}-\int_{\mathcal{X}}fd\mathbb{Q}\right|, \tag{2.9}$$

donde $\mathcal{F}$ es la clase de funciones acotadas Borel-medibles de valor real sobre $\mathcal{X}$.

Generalmente, estas medidas de distancia no son fáciles de calcular. Recientemente, [14] y [44] consideran la $\sigma-$álgebra $\mathcal{F}$ como la bola unidad en el espacio de Hilbert con kernel reproductivo, es decir $\mathcal{F}=\{f:\|f\|_{\mathcal{H}}\leq 1\}$ y definen una métrica $\gamma_k^2(\mathbb{P},\mathbb{Q})$ entre distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$ através de un kernel característico $k(x,x')$ como:

Definición 2.4.2. *Sean X y Y variables aleatorias definidas sobre un espacio topológico $\mathcal{X}$, con distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$ respectivamente y $\mathcal{F}=\{k:\mathcal{X}\mapsto\mathbb{R}:\|k\|_{\mathcal{H}}\leq 1\}$. Se define la medida de distancia entre $\mathbb{P}$ y $\mathbb{Q}$ como:*

$$\gamma_k(\mathbb{P},\mathbb{Q}) \;=\; \sup_{\|k\|_{\mathcal{H}}\leq 1}\left|\int_{\mathcal{X}}k_x d\mathbb{P}-\int_{\mathcal{X}}k_y d\mathbb{Q}\right|. \tag{2.10}$$

En el siguiente teorema se presenta una métrica entre distribuciones de probablidad en un RKHS.

Teorema 2.4.3. *Si las distribuciones de probabilidad $\mathbb{P}(x)$ y $\mathbb{Q}(y)$ admiten funciones de densidad $p(x)$ y $q(y)$ respectivamente con $d\mathbb{P}(x)=p(x)dx$ y $d\mathbb{Q}(y)=q(y)dy$, entonces $\gamma_k^2(\mathbb{P},\mathbb{Q})$ puede escribirse como*

$$\begin{aligned}\gamma_k^2(\mathbb{P},\mathbb{Q}) &= \left\|\int_{\mathcal{X}}k_x d\mathbb{P}-\int_{\mathcal{X}}k_y d\mathbb{Q}\right\|_{\mathcal{H}}^2=\int_{\mathcal{X}}\int_{\mathcal{X}}k(x,z)\,p(x)\,p(z)\,dxdz\\ &+ \int_{\mathcal{X}}\int_{\mathcal{X}}k(z,y)\,q(z)\,q(y)\,dzdy-2\int_{\mathcal{X}}\int_{\mathcal{X}}k(x,y)p(x)q(y)dxdy.\end{aligned} \tag{2.11}$$

La prueba se puede ver en [49].

Nótese que, el operador que hace el embebimiento en esta distancia está dado por

$$\mu_X=\mathbb{E}_X[k_X] \;=\; \int_{\mathcal{X}}k_x d\mathbb{P}(x)=\int_{\mathcal{X}}k_x p(x)\,dx. \tag{2.12}$$

Una de las pocas métricas entre distribuciones de probabilidad en un RKHS que se conocen en la literatura es la MMD, esta métrica se propone en [15] y es presentada en el siguiente teorema:

Teorema 2.4.4. *Supongamos que se tienen dos muestras aleatorias $\{\boldsymbol{x}_i\}_{i=1}^{N_x}$ y $\{\boldsymbol{y}_j\}_{j=1}^{N_y}$ tomadas de distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$ respectivamente donde $\boldsymbol{x}_i,\boldsymbol{x}\in\mathbb{R}^D$, y*

sea k un kernel característico. Si

$$\mathbb{P} = \sum_{i=1}^{N_x} k_{\boldsymbol{x}_i} p(\boldsymbol{x}_i) = \frac{1}{N_x} \sum_{i=1}^{N_x} k_{\boldsymbol{x}_i} \quad con \quad p(\boldsymbol{x}) = \frac{1}{N_x} \sum_{i=1}^{N_x} \delta(\boldsymbol{x} - \boldsymbol{x}_i), \tag{2.13}$$

$$\mathbb{Q} = \sum_{j=1}^{N_y} k_{\boldsymbol{y}_j} p(\boldsymbol{y}_j) = \frac{1}{N_y} \sum_{j=1}^{N_y} k_{\boldsymbol{y}_j} \quad con \quad p(\boldsymbol{y}) = \frac{1}{N_y} \sum_{j=1}^{N_y} \delta(\boldsymbol{y} - \boldsymbol{y}_j), \tag{2.14}$$

donde $\delta(\cdot)$ es la función delta Dirac, entonces

$$\begin{aligned} \gamma_k^2(\mathbb{P}, \mathbb{Q})_{MMD} &= \frac{1}{N_x^2} \sum_{j=1}^{N_x} \sum_{i=1}^{N_x} k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \frac{1}{N_y^2} \sum_{j=1}^{N_y} \sum_{i=1}^{N_y} k(\boldsymbol{y}_i, \boldsymbol{y}_j) \\ &- \frac{2}{N_x N_y} \sum_{j=1}^{N_y} \sum_{i=1}^{N_x} k(\boldsymbol{x}_i, \boldsymbol{y}_j). \end{aligned} \tag{2.15}$$

Note que si $p(x) = \frac{1}{N_x} \sum_{i=1}^{N_x} \delta(\mathbf{x} - \mathbf{x}_i)$, entonces el estimador de $\mathbb{E}_x[k_X]$ es dado por $\widehat{\mathbb{E}}_X[k_X] = \sum_{i=1}^{N_x} f(\mathbf{x}_i)$, donde $k_{\mathbf{x}} = f(\mathbf{x})$.

Teorema 2.4.5. *La métrica $\gamma_k^2(\mathbb{P}, \mathbb{Q})_{MMD}$ es un estimador consistente e insesgado de $\gamma_k^2(\mathbb{P}, \mathbb{Q})$.*

La prueba se puede ver en [14].

2.5 Resumen y comentarios del Capítulo 2

En este capítulo, se presentaron dos conceptos teóricos basados en el método embebimiento de distribuciones de probabilidad en un RKHS. El primer concepto fue la métrica entre las distribuciones de probabilidad en un RKHS. Este concepto nos permitirá construir en el Capítulo 3 nuevas métricas entre distribuciones de probabilidad y entre HMMs. El segundo concepto fue el operador de covarianza cruzada. Este concepto permitirá medir la relación de dependencia entre variables aleatorias. Usaremos este concepto en el Capítulo 4 para hacer predicción a corto plazo en un proceso autorregresivo.

Capítulo 3

Métricas entre modelos ocultos de Markov basadas en el método embebimiento de distribuciones de probabilidad en un RKHS

Las medidas de distancia entre HMMs han encontrado muchas aplicaciones en la teoría de la probabilidad y estadística y en la teoría de la información [19, 27]. Especialmente, en problemas de clasificación de series de tiempo se han propuesto muchas medidas de similitud como por ejemplo, las formas generalizadas de la distancia Euclidiana entre las matrices de probabilidad de las observaciones [21], la distancia de Bhattacharyya [2], la medida de distancia basada en el error de la probabilidad de Bayes [50], la medida de similaridad entre HMMs basada en la métrica de Wasserstein presentada en [8] y la divergencia de KL y sus modificaciones [57]. Sin embargo, algunas medidas de distancias como la divergencia KL y la medida de Bhattacharyya no son verdaderas medidas de distancia, dado que no satisfacen las propiedades de simetría y la desigualdad triangular respectivamente, por lo tanto estas medidas pueden producir bajo rendimiento en problemas de clasificación de series de tiempo [11, 58].

Los autores en [9] y en [56] proponen verdaderas medidas de distancia entre HMMs. Una desventaja que tiene la medida de distancia para HMMs propuesta en [56] llamada medida de distancia para modelos ocultos de Markov estacionarios (HSD), es que se define para distribuciones de probabilidad donde el espacio de entrada es de una dimensión y su cálculo en espacios de alta dimensión puede ser complicado, es decir, las integrales que aparecen en esta métrica son difíciles de resolver analíticamente.

El método embebimiento de distribuciones de probabilidad en un RKHS también ha sido usado con éxito para desarrollar medidas de distancia entre distribuciones de probabilidad [15] y entre procesos aleatorios con estructura Markoviana [9]. En la literatura de los embebimientos de distribuciones de probabilidad, una de las pocas métricas entre procesos aleatorios que se conoce es la que se presenta en [9], esta métrica es una combinación de la métrica MMD propuesta en [14] y el método de remuestreo de bootstrap.

En este capítulo se desarrollan dos métricas entre HMMs usando el método embebimiento de distribuciones de probabilidad en un RKHS, las cuales se utilizan para hacer clasificación de series de tiempo. Pero antes de desarrollar las métricas entre HMM, primero desarrollamos dos métricas entre distribuciones de probabilidad usando el método embebimiento junto con el estimador de Parzen.

3.1 Métricas entre distribuciones de probabilidad basadas en los RKHS

En esta sección, desarrollamos dos métricas entre las distribuciones de probabilidad en un RKHS. La diferencia entre las métricas que proponemos con la métrica MMD es que suponemos que las funciones de densidad de sus correspondientes distribuciones de probabilidad son funciones Gaussianas y que se pueden estimar mediante el estimador de Parzen. La dificultad de calcular métricas entre distribuciones de probabilidad usando los embebimientos cuando las densidades son continuas, es que las integrales que aparecen son difíciles de resolver analíticamente. También asumiremos que los kernels que permiten los embebimientos son el kernel Gaussiano y el kernel Laplaciano. Estas suposiciones sobre los kernels y sobre las funciones de densidad, nos permitirán encontrar dos nuevas métricas entre distribuciones de probabilidad de forma analítica.

3.1.1 Estimador de Parzen

El estimador de Parzen, es probablemente el estimador más popular y utilizado frecuentemente para la estimación no paramétrica de una función de densidad de probabilidad. La estimación de densidad de Parzen se refiere al problema de estimar la función de densidad de probabilidad $p(\mathbf{x})$ basada en una muestra aleatoria $\{\mathbf{x}_i\}_{i=1}^{N_x}$ de tamaño N_x. El estimador de Parzen para la función de densidad de probabilidad p viene dado por

$$\widehat{p}(\mathbf{x}) = \frac{1}{N_x}\sum_{i=1}^{N_x} k\left(\frac{\mathbf{x}-\mathbf{x}_i}{\sigma_p}\right), \quad \mathbf{x}_i,\ \mathbf{x}\in\mathbb{R}^D, \tag{3.1}$$

donde k es un kernel, que satisface ciertas condiciones de regularidad, generalmente es una función de densidad simétrica como por ejemplo la distribución normal, D es la dimensión del espacio de entrada y $\sigma_p > 0$ es un parámetro de suavización o bandwith. El estimador de Parzen ha sido muy utilizado en métodos de suavización por sus buenas propiedades estadísticas [40]. Si $k(\mathbf{x},\mathbf{x}')$ es el kernel Gaussiano, entonces el estimador de Parzen para $p(\mathbf{x})$, viene dado por

$$\widehat{p}(\mathbf{x}) = \frac{1}{N_x}\sum_{i=1}^{N_x} \frac{1}{\left(2\pi\sigma_p^2\right)^{D/2}} \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma_p^2}\right).$$

Una de las contribuciones de este trabajo de investigación es el desarrollo de una métrica entre distribuciones de probabilidad usando el método de embebimiento de distribuciones de probabilidad en un RKHS y el estimador de Parzen. La métrica que presentamos acontinuación se puede ver en [59], y viene dada por:

Teorema 3.1.1. *Si el kernel $k(\boldsymbol{x}, \boldsymbol{x}')$ es un kernel Gaussiano con parámetro $\boldsymbol{\Sigma}$, y los estimadores $\widehat{p}(\boldsymbol{x})$, y $\widehat{q}(\boldsymbol{y})$ son estimados como*

$$\widehat{p}(\boldsymbol{x}) = \frac{1}{N_x}\sum_{i=1}^{N_x} \frac{1}{\left(2\pi\sigma_p^2\right)^{D/2}} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\sigma_p^2}\right) = \frac{1}{N_x}\sum_{i=1}^{N_x} \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}_i, \boldsymbol{\Sigma}_p), \tag{3.2}$$

$$\widehat{q}(\boldsymbol{y}) = \frac{1}{N_y}\sum_{j=1}^{N_y} \frac{1}{\left(2\pi\sigma_q^2\right)^{D/2}} \exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{y}_j\|^2}{2\sigma_q^2}\right) = \frac{1}{N_y}\sum_{j=1}^{N_y} \mathcal{N}(\boldsymbol{y}|\boldsymbol{y}_j, \boldsymbol{\Sigma}_q), \tag{3.3}$$

respectivamente, donde $\boldsymbol{\Sigma}_p = \mathbf{I}\sigma_p^2$, $\boldsymbol{\Sigma}_q = \mathbf{I}\sigma_q^2$, $\mathbf{I}$ es la matriz identidad, σ_p y σ_q son los anchos de banda de los kernels, y D es la dimensión del espacio de entrada, entonces

$$\begin{aligned}\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q}) = &\frac{1}{N_x^2}\sum_{i,j=1}^{N_x} \hat{k}\left(\boldsymbol{x}_i, \boldsymbol{x}_j; 2\boldsymbol{\Sigma}_p\right) + \frac{1}{N_y^2}\sum_{i,j=1}^{N_y} \hat{k}\left(\boldsymbol{y}_i, \boldsymbol{y}_j; 2\boldsymbol{\Sigma}_q\right) \\ &- \frac{2}{N_x N_y}\sum_{i,j=1}^{N_x,N_y} \hat{k}\left(\boldsymbol{x}_i, \boldsymbol{y}_j; \boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_q\right),\end{aligned} \tag{3.4}$$

donde

$$\hat{k}\left(\boldsymbol{x}, \boldsymbol{x}'; \mathbf{S}\right) = \frac{|\boldsymbol{\Sigma}|^{1/2}}{|\boldsymbol{\Sigma}+\mathbf{S}|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}')^\top (\boldsymbol{\Sigma}+\mathbf{S})^{-1}(\boldsymbol{x}-\boldsymbol{x}')}{2}\right).$$

Este teorema es probado en el Apéndice A. Esta métrica la llamaremos métrica de Parzen y la denotamos por $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen}$.

Una propiedad deseable de un estimador es la consistencia. Los siguientes resultados permitirán probar que $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen}$ es un estimador consistente de $\gamma_k^2(\mathbb{P}, \mathbb{Q})$.

Proposición 3.1.2. *Sea*

$$\mathbb{E}_X[k_X] = \int_{\mathcal{X}} k_{\boldsymbol{x}} p(\boldsymbol{x}) d\boldsymbol{x},$$

el operador de embebimiento en un RKHS con $0 \leq k(\boldsymbol{x}, \boldsymbol{y}) \leq 1$ y $\widehat{p}(\boldsymbol{x}) = \frac{1}{N_x}\sum_{i=1}^{N_x} \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}_i, \boldsymbol{\Sigma}_p)$ el estimador de Parzen, entonces

$$\widehat{\mathbb{E}}_X[k_X] = \frac{1}{N_x}\sum_{i=1}^{N_x} g(\boldsymbol{x}_i), \tag{3.5}$$

es un estimador $\mathbb{E}_X[k_X]$ *donde* $g(\boldsymbol{x}_i) = \int_{\mathcal{X}} k_{\boldsymbol{x}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}_i \boldsymbol{\Sigma}_p) d\boldsymbol{x}$ *con* $\|g(\boldsymbol{x}_i)\| \leq 1$.

Prueba

$$\begin{aligned} \widehat{\mathbb{E}}_X[k_X] &= \int_{\mathcal{X}} k_{\mathbf{x}} \frac{1}{N_x} \sum_{i=1}^{N_x} \mathcal{N}(\mathbf{x}|\mathbf{x}_i, \boldsymbol{\Sigma}_p) d\mathbf{x} \\ &= \frac{1}{N_x} \sum_{i=1}^{N_x} \int_{\mathcal{X}} k_{\mathbf{x}} \mathcal{N}(\mathbf{x}|\mathbf{x}_i, \boldsymbol{\Sigma}_p) d\mathbf{x} = \frac{1}{N_x} \sum_{i=1}^{N_x} g(\mathbf{x}_i). \end{aligned} \tag{3.6}$$

Mostremos que $\|g(\mathbf{x}_i)\| \leq 1$.

$$\begin{aligned} \|g(\mathbf{x}_i)\|^2 &= \langle g(\mathbf{x}_i), g(\mathbf{x}_i) \rangle = \left\langle \int_{\mathcal{X}} k_{\mathbf{x}} \mathcal{N}(\mathbf{x}|\mathbf{x}_i, \boldsymbol{\Sigma}_p) d\mathbf{x}, \int_{\mathcal{X}} k_{\mathbf{y}} \mathcal{N}(\mathbf{y}|\mathbf{y}_i \boldsymbol{\Sigma}_q) d\mathbf{y} \right\rangle \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y}) \mathcal{N}(\mathbf{x}|\mathbf{x}_i, \boldsymbol{\Sigma}_p) \mathcal{N}(\mathbf{y}|\mathbf{y}_i \boldsymbol{\Sigma}_q) d\mathbf{x} d\mathbf{y}. \end{aligned} \tag{3.7}$$

Dado que $0 \leq k(\mathbf{x}, \mathbf{y}) \leq 1$ y $\int_{\mathcal{X}} \mathcal{N}(\mathbf{x}|\mathbf{x}_i, \boldsymbol{\Sigma}_p) = \int_{\mathcal{X}} \mathcal{N}(\mathbf{y}|\mathbf{y}_i, \boldsymbol{\Sigma}_q) = 1$, entonces $\|g(\mathbf{x}_i)\| \leq 1$. □

Proposición 3.1.3. $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen}$ *es un estimador consistente de* $\gamma_k^2(\mathbb{P}, \mathbb{Q})$.

Prueba Por la Proposición 3.1.2 $\|g(x_i)\| \leq 1$ y por el Teorema 2.4.5 se concluye que $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen}$ es un estimador consistente de $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})$. □

En la siguiente observación, nosotros presentamos algunos resultados sobre el estimador $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen}$.

Observación 3.1.4. *Las siguientes observaciones, muestran la relación de dependencia entre la métrica* $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen}$ *presentada en el Teorema* (3.1.1) *y la métrica* $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{MMD}$ *definida en* (2.15).

1. *Si* $\boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_q$, *entonces*

$$\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen} = \frac{|\boldsymbol{\Sigma}|^{1/2}}{|\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}_p|^{1/2}} \widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{MMD} = \lambda \widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{MMD}, \tag{3.8}$$

donde $0 \leq \lambda \leq 1$. *Esta observación muestra que la métrica* $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen}$ *se puede calcular a partir de la métrica* $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{MMD}$. *Es decir, multiplicando la metrica* $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{MMD}$ *por una constante* λ, *cuando las matrices de covarianza de las distribuciones de probabilidad son iguales.*

2. *Si* $\sigma_p,\ \sigma_q \to 0$, *entonces* $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen} = \widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{MMD}$. *Esto significa que el estimador* $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{MMD}$ *es un caso particular de el estimador* $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen}$.

3. Si $\boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_q$ *and* $\boldsymbol{\Sigma} = \lambda(\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}_p)$ *para* $0 \leq \lambda \leq 1$*, entonces* $\boldsymbol{\Sigma}_p = \left(\frac{1-\lambda}{2\lambda}\right)\boldsymbol{\Sigma}$*. Por lo tanto el estimador* $\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen}$ *es dado por*

$$\begin{aligned}\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{Parzen} &= \frac{\lambda^{D/2}}{N_x^2}\sum_{j=1}^{N_x}\sum_{i=1}^{N_x} k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \frac{\lambda^{D/2}}{N_y^2}\sum_{j=1}^{N_y}\sum_{i=1}^{N_y} k(\boldsymbol{y}_i, \boldsymbol{y}_j) \\ &- 2\frac{\lambda^{D/2}}{N_x N_y}\sum_{j=1}^{N_y}\sum_{i=1}^{N_x} k(\boldsymbol{x}_i, \boldsymbol{y}_j) = \lambda^{D/2}\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q})_{MMD}.\end{aligned}$$

3.1.2 Métrica entre distribuciones de probabilidad basada en distribuciones normales y en el kernel Laplaciano

En esta sección, desarrollamos una métrica entre las distribuciones de probabilidad basadas en el método de embebimiento entre distribuciones de probabilidad en un RKHS, suponiendo que el kernel del embebimiento es un kernel Laplaciano y las distribuciones de probabilidad son normales. Adiferencia de la métrica basada en el estimador de Parzen, la cual esta definida para espacios de entrada de alta dimensión, la métrica basada en el kernel Laplaciano que desarrollamos en esta sección estará definida para un espacio de entrada de una dimensión. Una de las razones de no poder definir la métrica basada en el kernel Laplaciano para espacios de entrada de alta dimensión, es por que la norma que define el kernel Laplaciano no es generada por un producto interno, por lo tanto no podemos resolver de forma cerrada, las integrales que definen esta métrica.

Teorema 3.1.5. *Si el kernel* $k(x, x'; \ell)$ *es un kernel Laplaciano donde* ℓ *es el parámetro de suavizado (ancho de banda) y*

$$\widehat{p}(x) = \frac{1}{\sigma_p\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right) \quad y \quad \widehat{q}(y) = \frac{1}{\sigma_q\sqrt{2\pi}}\exp\left(-\frac{(y-\mu_q)^2}{2\sigma_q^2}\right),$$

son estimadores de $p(x)$ *y* $q(y)$ *respectivamente, donde los parámetros* ℓ*,* μ_p*,* μ_q*,* σ_p*,* $\sigma_q \in \mathbb{R}$ *y los valores de entrada* $x, y \in \mathbb{R}$*, entonces un nuevo estimador de la métrica entre las distribuciones de probabilidad* $\mathbb{P}$ *y* $\mathbb{Q}$ *es obtenido a partir de la expresión* (2.11)

$$\begin{aligned}\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q}) &= \frac{1}{2\pi\sigma_p^2}\left(\mathcal{I}_1(\mu_p, \sigma_p, \ell) + \mathcal{I}_2(\mu_p, \sigma_p, \ell)\right) \\ &+ \frac{1}{2\pi\sigma_q^2}\left(\mathcal{I}_1(\mu_q, \sigma_q, \ell) + \mathcal{I}_2(\mu_q, \sigma_q, \ell)\right) \\ &- \frac{1}{\pi\sigma_p\sigma_q}\left(\mathcal{I}_1(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell) + \mathcal{I}_2(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell)\right), \qquad (3.9)\end{aligned}$$

donde

$$\mathcal{I}_1(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell) = f_1(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell)\frac{\sigma_p\sqrt{\pi}}{\sigma_q}\left(1 - \operatorname{erf}\left(\frac{d_1(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell)\,\sigma_p}{2\sigma_q\sqrt{\sigma_q^2 + \sigma_p^2}}\right)\right), \quad (3.10)$$

$$d_1 = d_1\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right) = \frac{\sigma_q}{\sqrt{2}\ell^2\sigma_p^2}\left(2\ell^2(\mu_q - \mu_p) + \sigma_p^2 + \sigma_q^2\right), \tag{3.11}$$

$$\begin{aligned} f_1\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right) &= \sigma_q^2\sqrt{\pi}\exp\left(\frac{\left(2\ell^2\mu_q + \sigma_q^2\right)^2}{8\ell^4\sigma_q^2}\left(1 - \frac{\sigma_q^2}{\sigma_p^2}\right)\right) \\ &\times \exp\left(-\frac{\left(2\ell^2\mu_q + \sigma_q^2\right)}{2\ell^2}\left(\frac{1}{2\ell^2} - \frac{\mu_p}{\sigma_p}\right) - \frac{1}{2}\left(\frac{\mu_q^2}{\sigma_q^2} + \frac{\mu_p^2}{\sigma_p^2}\right) + \frac{\sigma_p^2 d_1^2}{4\sigma_q^2}\right), \end{aligned} \tag{3.12}$$

$$\begin{aligned} \mathcal{I}_2\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right) &= f_2\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right) \\ &\times \frac{\sigma_q\sqrt{\pi}}{\sigma_p}\left(1 - \operatorname{erf}\left(\frac{d_2\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right)\sigma_q^2}{2\sigma_p\sqrt{\sigma_p^2 + \sigma_q^2}}\right)\right) \end{aligned} \tag{3.13}$$

$$d_2 = d_2\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right) = \frac{\sigma_p}{\sqrt{2}\ell^2\sigma_q^2}\left(2\ell^2\left(\mu_p - \mu_q\right) + \sigma_q^2 + \sigma_p^2\right), \tag{3.14}$$

$$\begin{aligned} f_2\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right) &= \sigma_p^2\sqrt{\pi}\exp\left(\frac{\left(2\ell^2\mu_p + \sigma_p^2\right)^2}{8\ell^4\sigma_p^2}\left(1 - \frac{\sigma_p^2}{\sigma_q^2}\right)\right) \\ &\times \exp\left(-\frac{\left(2\ell^2\mu_p + \sigma_p^2\right)}{2\ell^2}\left(\frac{1}{2\ell^2} - \frac{\mu_q}{\sigma_q^2}\right) - \frac{1}{2}\left(\frac{\mu_p^2}{\sigma_p^2} + \frac{\mu_q^2}{\sigma_q^2}\right) + \frac{\sigma_q^2 d_2^2}{4\sigma_p^2}\right). \end{aligned} \tag{3.15}$$

Este teorema es probado en el Apéndice A. Esta métrica la llamaremos métrica Laplaciana y la denotamos por $\widehat{\gamma_k^2}\left(\mathbb{P}, \mathbb{Q}\right)_{Laplaciana}$.

La generalización del Teorema 3.1.5 se presenta en el siguiente corolario.

Corolario 3.1.6. *Si $k(\boldsymbol{x}, \boldsymbol{y}; \ell)$ es un kernel Laplaciano de dimensión n, es decir*

$$k(\boldsymbol{x}, \boldsymbol{y}; \ell) = \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{y}\|_1}{2\ell^2}\right), \quad \textit{donde} \quad \boldsymbol{x},\ \boldsymbol{y} \in \mathbb{R}^n,$$

y

$$\widehat{p}(\boldsymbol{x}) = \frac{1}{|\boldsymbol{\Sigma}_p|^{1/2}(2\pi)^{n/2}}\exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_p^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_p)}{2}\right) \quad \textit{donde} \quad \boldsymbol{\Sigma}_p = diag\left(\sigma_{p1}^2, \sigma_{p2}^2, \cdots, \sigma_{pn}^2\right),$$

$$\widehat{q}(\boldsymbol{y}) = \frac{1}{|\boldsymbol{\Sigma}_q|^{1/2}(2\pi)^{n/2}}\exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{\mu}_q)^\top \boldsymbol{\Sigma}_q^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_q)}{2}\right) \quad \textit{donde} \quad \boldsymbol{\Sigma}_q = diag\left(\sigma_{q1}^2, \sigma_{q2}^2, \cdots, \sigma_{qn}^2\right),$$

son estimadores de $p(x)$ y $q(y)$ respectivamente, donde los parámetros $\ell \in \mathbb{R}$, $\boldsymbol{\mu}_p = (\mu_{p1}, \mu_{p2}, \cdots, \mu_{pn})$, $\boldsymbol{\mu}_q = (\mu_{q1}, \mu_{q2}, \cdots, \mu_{qn}) \in \mathbb{R}^n$ y $\boldsymbol{\Sigma}_p$, $\boldsymbol{\Sigma}_q \in \mathbb{R}^{n\times n}$ son matrices diagonales. Entonces el estimador de la métrica $\gamma_k^2(\mathbb{P}, \mathbb{Q})$ entre las distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$ viene dado por la expresión

$$\begin{aligned}\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q}) &= \prod_{i=1}^{n} \frac{1}{2\pi\sigma_{pi}^2} \left(\mathcal{I}_1(\mu_{pi}, \sigma_{pi}, \ell) + \mathcal{I}_2(\mu_{pi}, \sigma_{pi}, \ell)\right) \\ &+ \prod_{i=1}^{n} \frac{1}{2\pi\sigma_{qi}^2} \left(\mathcal{I}_1(\mu_{qi}, \sigma_{qi}, \ell) + \mathcal{I}_2(\mu_{qi}, \sigma_{qi}, \ell)\right) \\ &- \prod_{i=1}^{n} \frac{1}{\pi\sigma_{pi}\sigma_{qi}} \left(\mathcal{I}_1(\mu_{pi}, \mu_{qi}, \sigma_{pi}, \sigma_{qi}, \ell) + \mathcal{I}_2(\mu_{pi}, \mu_{qi}, \sigma_{pi}, \sigma_{qi}, \ell)\right), \quad (3.16)\end{aligned}$$

donde las funciones $\mathcal{I}_1$ y $\mathcal{I}_2$ son definidas como en las ecuaciones 3.10 y 3.13 respectivamente.

En la siguiente sección presentamos algunos resultados que permiten generar métricas a partir de una métrica conocida.

3.1.3 Funciones que preservan métricas

En esta sección, presentamos algunos resultados (ver [10]) y ejemplos específicos de funciones que son generadoras de métricas.

Proposición 3.1.7. *Sea $d : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}^+$ una métrica y $f : [0, \infty) \longrightarrow [0, \infty)$ una función estrictamente creciente y continua tal que*

(*a*) $f^{-1}(0) = \{0\}$.

(*b*) $f(x + y) \leq f(x) + f(y)$.

Entonces $d^* = f(d)$ es una métrica.

Prueba Mostremos que $d^* = f(d)$ es una métrica. Verifiquemos la desigualdad triangular, las otras dos propiedades de métricas son fáciles de verificar. Sean $x, y, z \in \mathcal{X}$ y $a = d(x, z)$, $b = d(y, z)$ y $c = d(x, y)$. Como d es una métrica, entonces

$$d(x, y) \leq d(x, z) + d(y, z). \quad (3.17)$$

Aplicando f a la ecuación 3.17 y dado que f cumple la propiedad (b), entonces obtenemos

$$f(d(x, y)) \leq f(d(x, z) + d(y, z)) \leq f(d(x, z)) + f(d(y, z)),$$

luego

$$d^*(c) \leq d^*(a) + d^*(b).$$

Esto implica que d^* satisface la desigualdad triangular, luego d^* es una métrica.□

Con base en de la Proposición 3.1.7 se sigue que $d_\gamma(\mathbb{P}, \mathbb{Q})_{Parzen} = f(\gamma_k(\mathbb{P}, \mathbb{Q})_{Parzen})$ y $d_\gamma(\mathbb{P}, \mathbb{Q})_{Laplaciana} = f(\gamma_k(\mathbb{P}, \mathbb{Q})_{Laplaciana})$ son métricas si f satisface las condiciones (a)

y (b) de la proposición 3.1.7.

El siguiente resultado, permite construir funciones que preservan métricas.

Proposición 3.1.8. *Si $g : [0,\infty) \longrightarrow [0,\infty)$ es una función decreciente y continua, entonces*

$$f(x) = \int_0^x g(t)dt,$$

satisface las condiciones 1 y 2 de la Proposición (3.1.7).

La prueba se puede ver en [53].

A partir de la Proposición 3.1.8, presentamos algunos resultados sobre métricas entre distribuciones de probabilidad que son generadas por una métrica conocida. Estos resultados no solo son resultados teóricos interesantes, sino que permiten construir métricas a partir de una métrica conocida con el fin de mejorar el rendimiento en clasificación de series de tiempo.

Ejemplo 3.1.9. *Presentamos algunos ejemplos de métricas generadas por las métricas $\gamma_k(\mathbb{P},\mathbb{Q})_{Parzen}$ y $\gamma_k(\mathbb{P},\mathbb{Q})_{Laplaciana}$.*

1. *Si*

$$g(t) = \frac{1}{1+t}, \quad entonces \quad f(x) = \int_0^x \frac{dt}{1+t} = \ln(1+x) \quad x \in [0,\infty).$$

Por lo tanto

$$d_\gamma(\mathbb{P},\mathbb{Q})_{Parzen} = \ln\left(1+\gamma_k(\mathbb{P},\mathbb{Q})_{Parzen}\right),$$

y

$$d_\gamma(\mathbb{P},\mathbb{Q})_{Laplaciana} = \ln\left(1+\gamma_k(\mathbb{P},\mathbb{Q})_{Laplaciana}\right),$$

son métricas, donde $\gamma_k(\mathbb{P},\mathbb{Q})_{Parzen}, \gamma_k(\mathbb{P},\mathbb{Q})_{Laplaciana} \in [0,\infty)$.

2. *Si*

$$g(t) = \frac{1}{(1+t)^2}, \quad entonces \quad f(x) = \int_0^x \frac{dt}{(1+t)^2} = \frac{x}{1+x} \quad x \in [0,\infty).$$

Por consiguiente

$$d_\gamma(\mathbb{P},\mathbb{Q})_{Parzen} = \frac{\gamma_k(\mathbb{P},\mathbb{Q})_{Parzen}}{1+\gamma_k(\mathbb{P},\mathbb{Q})_{Parzen}},$$

y

$$d_\gamma(\mathbb{P},\mathbb{Q})_{Laplaciana} = \frac{\gamma_k(\mathbb{P},\mathbb{Q})_{Laplaciana}}{1+\gamma_k(\mathbb{P},\mathbb{Q})_{Laplaciana}},$$

son métricas, donde $\gamma_k(\mathbb{P},\mathbb{Q})_{Laplaciana}, \gamma_k(\mathbb{P},\mathbb{Q})_{Parzen} \in [0,\infty)$.

3. Si

$$g(t) = \frac{1}{1+t^2}, \quad \textit{entonces} \quad f(x) = \int_0^x \frac{dt}{1+t^2} = \arctan(x) \quad x \in [0, \tfrac{\pi}{2}).$$

En consecuencia

$$d_\gamma(\mathbb{P},\mathbb{Q})_{Parzen} = \arctan\left(\gamma_k(\mathbb{P},\mathbb{Q})_{Parzen}\right)$$

y

$$d_\gamma(\mathbb{P},\mathbb{Q})_{Laplaciana} = \arctan\left(\gamma_k(\mathbb{P},\mathbb{Q})_{Laplaciana}\right)$$

son métricas, donde $\gamma_k(\mathbb{P},\mathbb{Q})_{Parzen}, \gamma_k(\mathbb{P},\mathbb{Q})_{Laplaciana} \in [0, \frac{\pi}{2})$.

4. Si

$$g(t) = \exp(-t), \quad \textit{entonces} \quad f(x) = \int_0^x \frac{dt}{\exp(t)} = 1 - \exp(-x) \quad x \in [0, \infty).$$

Así pues

$$d_\gamma(\mathbb{P},\mathbb{Q})_{Parzen} = 1 - \exp(-\gamma_k\left(\mathbb{P},\mathbb{Q}\right)_{Parzen}),$$

y

$$d_\gamma(\mathbb{P},\mathbb{Q})_{Laplaciana} = 1 - \exp(-\gamma_k\left(\mathbb{P},\mathbb{Q}\right)_{Laplaciana}),$$

son métricas, donde $\gamma_k(\mathbb{P},\mathbb{Q})_{Parzen}, \gamma_k(\mathbb{P},\mathbb{Q})_{Laplaciana} \in [0, \infty)$.

3.2 Resultados experimentales para la métrica basada en los RKHS y el estimador de Parzen

En esta sección presentamos el rendimiento de la métrica basada en el método embebimiento de distribuciones de probabilidad en un RKHS y el estimador de Parzen (PKE). Para este propósito, nuestra métrica se evalúa en 24 de las treinta y una bases de datos binarias de la UCR Time Classification Archive [7], utilizando el clasificador K-vecino más cercano (KNN) para $K = 1, 3, 5$. La descripción de las bases de datos está en la Tabla 3.1, la cual muestra el tamaño de los datos de entrenamiento y los datos de prueba para cada una de las clases. Comparamos el rendimiento de la métrica que proponemos PKE con la métrica MMD y la métrica Euclideana (EUC).

La clasificación de series de tiempo, es probablemente el método más popular en el aprendizaje de máquinas. La clasificación asocia series de tiempo entre clases predefinidas. El problema de clasificación de series de tiempo se define de la siguiente manera: Dado un conjunto de series de tiempo (bases de datos) $\mathcal{D} = \{T_1, T_2, \cdots, T_N\}$ y un conjunto de clases $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_M\}$, el problema de clasificar estas series de tiempo, es definir una función $g : \mathcal{D} \longrightarrow \mathcal{C}$, donde a cada serie de tiempo T_i se le asigna

una clase. Además una clase $\mathcal{C}_j$ contiene a las series de tiempo asignadas en ella, es decir $\mathcal{C}_j = \{T_i : g(T_i) = \mathcal{C}_j, 1 \leq i \leq N, T_i \in \mathcal{D}\}$.

El algoritmo de clasificación KNN que utilizamos en este trabajo, asume algún conocimiento de los datos o realiza fases de entrenamiento para estas clasificaciones. Este algoritmo está basado en el uso de métricas, en nuestro caso métricas entre distribuciones de probabilidad, y asume que las series de tiempo o datos de entrenamiento son el modelo de los datos. Cuando se tiene que hacer una clasificación dado un nuevo conjunto de series de tiempo, se calcula su distancia con cada uno de los elementos del modelo. La serie de tiempo es asignada a la clase mayoritaria que contiene los k vecinos más cercanos a la serie de tiempo [48].

Tabla 3.1: Las treinta y una bases de datos binarias con el tamaño del conjunto de entrenamiento y el tamaño del conjunto de prueba, usadas en este trabajo para comparar el rendimiento de las métricas que proponemos.

Database	Train size	Test size
BeetleFly	10 10	10 10
BirdChicken	10 10	10 10
Coffee	14 14	15 13
Computers	125 125	125 125
DistalPhalanxOutlineCorrect	115 161	222 378
ECG200	31 69	36 64
ECGFiveDays	14 9	428 433
Earthquakes	104 35	264 58
FordA	681 639	1846 1755
FordB	401 409	1860 1776
GunPoint	24 26	76 74
Ham	52 57	51 54
HandOutlines	133 237	362 638
Herring	39 25	38 26
ItalyPowerDemand	34 33	513 516
Lighting2	20 40	28 33
MiddlePhalanxOutlineCorrect	125 166	212 388
MoteStrain	10 10	675 577
PhalangesOutlinesCorrect	628 1172	332 526
ProximalPhalanxOutlineCorrect	194 406	92 199
ShapeletSim	10 10	90 90
SonyAIBORobotSurface	6 14	343 258
SonyAIBORobotSurfaceII	11 16	365 588
Strawberry	132 238	219 394
ToeSegmentation1	20 20	120 108
ToeSegmentation2	18 18	106 24
TwoLeadECG	12 11	569 570
Wine	30 27	27 27
WormsTwoClass	33 44	76 105
Wafer	97 903	665 5499
Yoga	137 163	1393 1607

En los experimentos, se supone que los parámetros de suavización o bandwith de las distribuciones de probabilidad usando el estimador de Parzen, son estimados mediante un enfoque de optimización del ancho de banda del kernel [43], este método consiste en minimizar el error cuadrático medio integrado (MISE) entre la distribución de probabilidad estimada y la real, esto es

$$MISE(\sigma_p) = \mathbb{E}\left[\int \left(\widehat{p}(\mathbf{x}) - p(\mathbf{x})\right)^2 d\mathbf{x}\right].$$

Un concepto que utlizaremos en el análisis de los resultados experimentales, es el de coeficiente de variación (CV) multivariado. El CV multivariado es una medida de variabilidad de un conjunto de series de tiempo, este coeficiente se puede usar para comparar la variabilidad entre dos conjuntos de series de tiempo. El CV multivariado lo definimos como en [26].

Definición 3.2.1. *Considere un conjunto de N series de tiempo de longitud T, con vector media $\mu \neq 0$ y matriz de covarianza Σ, el CV es definido como*

$$CV = \left(\text{Trace}(\Sigma)\left(\mu^\top \mu\right)^{-1}\right)^{1/2}.$$

A continuación presentamos el rendimiento de la métrica PKE con respecto a las métricas MMD y EUC.

Tabla 3.2: Compara el rendimiento de la métricas PKE, MMD y EUC, usando el algoritmo KNN para $K = 1, 3, 5$

	K=1			K=3			K=5		
	PKE	*MMD*	*EUC*	*PKE*	*MMD*	*EUC*	*PKE*	*MMD*	*EUC*
BeetleFly	0.7500	0.6500	0.7500	0.7000	0.6000	0.6500	0.7500	0.6500	0.6000
BirdChicken	0.9500	0.7000	0.5500	0.8500	0.9000	0.4500	0.8500	0.6500	0.5500
Coffee	0.8214	0.6071	1.000	0.8214	0.6429	1.000	0.8571	0.5714	0.9643
DistalPhalanxOutlineCorrect	0.7417	0.5917	0.7517	0.7450	0.6300	0.7583	0.7750	0.6267	0.7633
ECG200	0.7800	0.6200	0.8800	0.7700	0.7300	0.9000	0.7800	0.7600	0.9000
ECGFiveDays	0.8641	0.7224	0.7967	0.8351	0.7480	0.7398	0.7851	0.7468	0.6121
Earthquakes	0.7142	0.8199	0.6739	0.7671	0.8199	0.7422	0.8106	0.8199	0.7857
FordA	0.5460	0.5318	0.6590	0.5590	0.5221	0.6715	0.5740	0.5149	0.6862
FordB	0.5294	0.5102	0.5578	0.5490	0.5195	0.5833	0.5503	0.5314	0.5833
GunPoint	0.8733	0.8200	0.9133	0.8467	0.8400	0.8733	0.7533	0.7533	0.8000
Ham	0.4857	0.4857	0.6000	0.5619	0.5714	0.5905	0.6476	0.6381	0.6286
Herring	0.6250	0.5938	0.5156	0.6094	0.5938	0.5625	0.5625	0.5938	0.5156
Lighting2	0.7377	0.7541	0.7541	0.7213	0.7049	0.7705	0.7213	0.7049	0.7213
MiddlePhalanxOutlineCorrect	0.6317	0.6317	0.7533	0.6617	0.6233	0.7717	0.6583	0.5617	0.7600
ProximalPhalanxOutlineCorrect	0.7595	0.6838	0.8076	0.8076	0.6838	0.8488	0.8007	0.6838	0.8419
ShapeletSim	0.5222	0.5500	0.5389	0.5056	0.4722	0.5278	0.5056	0.5222	0.5444
SonyAIBORobotSurface	0.7820	0.6738	0.6955	0.8236	0.6256	0.5740	0.8053	0.5757	0.4692
SonyAIBORobotSurfaceII	0.7618	0.6180	0.8593	0.7786	0.6180	0.7985	0.7775	0.6337	0.7712
Strawberry	0.8825	0.6835	0.9380	0.8842	0.7227	0.9233	0.8564	0.7113	0.9233
ToeSegmentation1	0.7105	0.7149	0.6798	0.7412	0.7149	0.6053	0.7193	0.7105	0.6140
ToeSegmentation2	0.8000	0.8154	0.8077	0.8000	0.8154	0.8231	0.7385	0.8154	0.8462
TwoLeadECG	0.8973	0.5443	0.7471	0.8850	0.5540	0.6348	0.8797	0.5505	0.5970
Wine	0.7593	0.6111	0.6111	0.7222	0.5741	0.5555	0.6852	0.5741	0.5370
WormsTwoClass	0.5801	0.5801	0.5856	0.5856	0.5811	0.5911	0.5414	0.5811	0.5967
Ganador	7/24	5/24	**14/24**	7/24	2/24	**15/24**	11/24	2/24	**12/24**

La Tabla 3.2 muestra el rendimiento de la métrica PKE sobre 24 bases de datos del conjunto de datos disponibles en la UCR. A partir de la Tabla 3.2, podemos observar que la métrica PKE tiene un mejor rendimiento que las métricas MMD y EUC en el 45.83% de las bases de datos UCR, usando el clasificador KNN para $K = 5$. También observamos que la métrica EUC tiene un mejor rendimiento en clasificación que las métricas PKE y MMD en el 58.33% y 62.5% de las bases de datos de la UCR para $K = 1, 3$ respectivamente. A partir de la Tabla 3.2, observamos que las bases de datos donde la métrica que proponemos PKE tiene mejor rendimiento, son bases de datos donde los conjuntos de entrenamiento y de prueba son muy parecidos o donde los datos de prueba son muy grandes comparados con los datos de entrenamiento, ejemplo de esto son las bases de datos BeetleFly, BirdChicken, ECGFiveDays, Herring, SonyAIBORobotSurface, TwoLeadECG y Wine. Esto se debe probablemente a que nuestra métrica es un estimador consistente, por consiguiente su rendimiento para hacer clasificación de series de tiempo, aumenta cuando estas series son muy grandes. De la Tabla 3.2 se puede observar que la métrica PKE tiene sus mejores rendimientos en las bases de conjuntos de datos donde el CV de los datos de entrenamiento y datos de prueba son pequeños como por ejemplo, las bases de datos Wine, DistalPhalanxOutlineCorrect y TwoLeadECG. La Tabla 3.2 muestra también, que la métrica PKE tiene un bajo rendimiento en clasificación de series de tiempo y es superada por la métrica EUC, cuando los conjuntos de entrenamiento y de prueba de las bases de datos tienen un CV muy grande, como por ejemplo las bases de datos ForA, ForB y WormsTwoClass. La principal razón puede ser porque la métrica PKE se basa en un kernel suave como el kernel Gaussiano y en distribuciones de probabilidad suaves, esto permite que la métrica PKE tenga un buen rendimiento en clasificación cuando las series de tiempo son suaves esto es, cuando las series de tiempo tienen un pequeño CV. La Tabla 3.3 compara

Tabla 3.3: Compara el rendimiento de la métricas PKE y MMD basadas en el método embebimiento de distribuciones de probabilidad en un RKHS, usando el algoritmo KNN para $K = 1, 3, 5$

	K=1		K=3		K=5	
	PKE	*MMD*	*PKE*	*MMD*	*PKE*	*MMD*
BeetleFly	0.7500	0.6500	0.7000	0.6000	0.7500	0.6500
BirdChicken	0.9500	0.7000	0.8500	0.9000	0.8500	0.6500
Coffee	0.8214	0.6071	0.8214	0.6429	0.8571	0.5714
DistalPhalanxOutlineCorrect	0.7417	0.5917	0.7450	0.6300	0.7750	0.6267
ECG200	0.7800	0.6200	0.7700	0.7300	0.7800	0.7600
ECGFiveDays	0.8641	0.7224	0.8351	0.7480	0.7851	0.7468
Earthquakes	0.7142	0.8199	0.7671	0.8199	0.8106	0.8199
FordA	0.5460	0.5318	0.5590	0.5221	0.5740	0.5149
FordB	0.5294	0.5102	0.5490	0.5195	0.5503	0.5314
GunPoint	0.8733	0.8200	0.8467	0.8400	0.7533	0.7533
Ham	0.4857	0.4857	0.5619	0.5714	0.6476	0.6381
Herring	0.6250	0.5938	0.6094	0.5938	0.5625	0.5938
Lighting2	0.7377	0.7541	0.7213	0.7049	0.7213	0.7049
MiddlePhalanxOutlineCorrect	0.6317	0.6317	0.6617	0.6233	0.6583	0.5617
ProximalPhalanxOutlineCorrect	0.7595	0.6838	0.8076	0.6838	0.8007	0.6838
ShapeletSim	0.5222	0.5500	0.5056	0.4722	0.5056	0.5222
SonyAIBORobotSurface	0.7820	0.6738	0.8236	0.6256	0.8053	0.5757
SonyAIBORobotSurfaceII	0.7618	0.6180	0.7786	0.6180	0.7775	0.6337
Strawberry	0.8825	0.6835	0.8842	0.7227	0.8564	0.7113
ToeSegmentation1	0.7105	0.7149	0.7412	0.7149	0.7193	0.7105
ToeSegmentation2	0.8000	0.8154	0.8000	0.8154	0.7385	0.8154
TwoLeadECG	0.8973	0.5443	0.8850	0.5540	0.8797	0.5505
Wine	0.7593	0.6111	0.7222	0.5741	0.6852	0.5741
WormsTwoClass	0.5801	0.5801	0.5856	0.5811	0.5414	0.5811
Ganador	**16/24**	5/24	**20/24**	4/24	**18/24**	5/24

el rendimiento de las métricas PEK y MMD basadas en el método embebimiento de distribuciones de probabilidad en un RKHS. De la Tabla 3.3 se puede concluir que la métrica PKE tiene mejor rendimiento en clasificación que la métrica MMD en el 66.67%, 83.33% y 75% de las bases de datos UCR para $K = 1, 3, 5$ respectivamente. Esto se debe probablemente a que la métrica PKE le da más relevancia a las muestras, estimando los pesos que determinan las funciones de probabilidad mediante el método MISE definido anteriormente. Otra posible razón del mejor rendimiento en clasificación de series de tiempo de la métrica PKE comparado con la métrica MMD, es debido a que la métrica PKE es un estimador más robusto que la métrica MMD [59].

Cabe resaltar que en [59] se introdujo por primera vez la métrica de Parzen con el propósito de comparar dos distribuciones de datos en un RKHS usando estimadores de densidad suaves. Esta métrica fue empleada en este trabajo con buen rendimiento en métodos de Aproximacion Bayesiana Computacional (ABC).

3.3 Modelos ocultos de Markov (HMMs)

Los HMMs son uno de los enfoques más exitosos en estadística aplicada para la modelización de datos secuenciales, para la predicción de series de tiempo y para el reconocimiento automático de patrones [5, 4]. Un HMM es un proceso estocástico que consiste en una secuencia de observaciones, la cual es tomada de un proceso de Markov subyacente de un número finito de estados discretos no observables conocido como estados ocultos [55]. En esta sección vamos a explicar un HMM para una secuencia de observación multidimensional.

Definición 3.3.1. *Sea* $(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_T)$ *una secuencia de observaciones multivariada de longitud* T *la cual es tomada de una secuencia de Markov no observable* $(H_1, H_2, \cdots, H_T)$. *Note que* $\mathbf{X}_t$ *y* H_t *para* $t = 1, 2, \cdots, T$ *toman valores en los conjuntos* $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M\}$ *y* $\{h_1, h_2, \cdots, h_N\}$ *respectivamente, con* N *el número de estados y* M *el número de observaciones. Aquí,* $\mathbf{x}_k \in \mathbb{R}^D$ *para* $k = 1, 2, \cdots, M$ *y* h_i *es el estado* i, *para* $i = 1, 2, \cdots, N$.

Un HMM esta completamente definido por la tripla $\rho = (\pi, A, B)$. *El primer término* $\pi \in \mathbb{R}^N$ *representa la distribución de estado inicial con* $\pi_i = p(H_1 = h_i)$.

La matriz de transición $A \in \mathbb{R}^{N \times N}$ *tiene entradas* $a_{ij} = p(H_t = h_j | H_{t-1} = h_i)$, $\forall i, j = 1, 2, \cdots, N$, *las cuales representa la probabilidad de transición del estado* h_i *al estado* h_j. *El vector de emisión* $B \in \mathbb{R}^{N \times M}$ *tiene entradas*

$$b_i(\mathbf{x}) = p(\mathbf{X}_t = \mathbf{x} | H_t = h_i),\ i = 1, 2, \cdots, N,\ k = 1, 2, \cdots, M,$$

y representa la probabilidad de emitir $\mathbf{x}$ *desde el estado* h_i.

De acuerdo con la regla del producto para distribuciones, la distribución conjunta del HMM se puede definir de la siguiente manera:

Definición 3.3.2. *Sea* $(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_T)$ *una secuencia de observaciones tal que se supone que sigue una distribución dada por el HMM definido anteriormente. La*

probabilidad de la secuencia $(\mathbf{X}_1,\ \mathbf{X}_2,\ \cdots,\ \mathbf{X}_T)$ *se define como:*

$$p\left(\mathbf{X}_1=\mathbf{x}_1,\ \cdots,\ \mathbf{X}_T=\mathbf{x}_T\right)=p\left(\mathbf{x}_1,\ \cdots,\ \mathbf{x}_T\right)=\sum_{i_1=1}^{N}\cdots\sum_{i_T=1}^{N}\pi_{i_1}b_{i_1}(\mathbf{x}_1)\cdots a_{i_{T-1},i_T}b_{i_T}(\mathbf{x}_T).$$

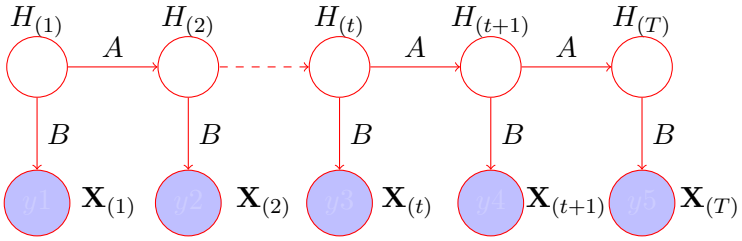


Figura 3.1: Diagrama de un HMM multivariado para T puntos de tiempo. Los escalares $H_{(t)}$ y $\mathbf{X}_{(t)}$ representan el estado oculto y los valores observados en instantánea t, respectivamente. Los términos A y B son la matriz de transición y el vector de emisión respectivamente.

En este trabajo de investigación estamos interesados en la distribución estacionaria asociada a un HMM. La distribución estacionaria de un vector de observaciones se define como:

Definición 3.3.3. *Sea* $\pi_{s,i}$ *una distribución estacionaria, es decir* $\pi_{t,i}=\pi_{t-1,i}=\pi_{s,i}$, $t\geq t_s$, *donde* t_s *es el tiempo de estacionariedad del HMM [56], entonces la distribución estacionaria del HMM puede ser escrita como*

$$p(\boldsymbol{x})=\sum_{i=1}^{N}\pi_{s,i}b_i(\boldsymbol{x}). \tag{3.18}$$

Supongamos que la probabilidad de emisión: $b_i(\mathbf{x})$ para $i=1,2,\cdots N$ es dada por un modelo de mezcla de Gaussianas (GMM) donde $M_{\mathbb{P}}$ es el número de componentes de la mezcla, es decir

$$b_i(\mathbf{x})=\sum_{j=1}^{M_{\mathbb{P}}}\alpha_{i,j}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_{i,j},\boldsymbol{\Sigma}_{i,j}\right),\ \text{con}\ \sum_{j=1}^{M_{\mathbb{P}}}\alpha_{i,j}=1\ \text{y}\ \alpha_{i,j}\geq 0,$$

donde $\alpha_{i,j}$ es el coeficiente para la j-ésima componente de la mezcla en el estado i. $\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_{i,j},\boldsymbol{\Sigma}_{i,j}\right)$ es una distribución Gaussiana donde $\mathbf{x}$ es un vector de observaciones, $\boldsymbol{\mu}_{i,j}$ es el vector media de la componente j del estado i y $\boldsymbol{\Sigma}_{i,j}$ es la matriz de covarianza de la componente j del estado i. Estos parámetros pueden ser estimados através del algoritmo Maximización de la Esperanza (EM) [31, 34].

3.4 Métricas entre HMMs estacionarios usando el método de embebimiento de distribuciones de probabilidad en un RKHS

En esta sección, desarrollamos dos nuevas métricas para discriminar dos HMMs usando la teoría de los embebimientos de distribuciones de probabilidad en RKHS. Supongamos

que las distribuciones de probabilidad $p(\cdot)$ y $p(\cdot)$, estan dadas por distribuciones estacionarias de las observaciones de los HMMs como se observa en la Ecuación (3.18) donde se asume que las emisiones de probabilidad son mezclas de Gaussianas. Note que la solución para la Ecuación (2.11) depende del tipo de kernel característico $k(\cdot,\cdot)$, que nosotros elegimos para el espacio de Hilbert encajado. En este trabajo, asumiremos un kernel Gaussiano y un kernel Laplaciano, con el propósito de poder calcular de forma analítica métricas basadas en los RKHS. Aunque debemos restringir la discusión para un kernel Gaussiana y un kernel Laplaciano, es posible asumir diferentes tipos de kernels característicos, obteniendo una amplia gama de métricas.

Definición 3.4.1. *Sean $p(\boldsymbol{x})$ y $q(\boldsymbol{y})$ distribucones estacionarias de dos HMMs con sus estimadores $\widehat{p}(\boldsymbol{x})$ y $\widehat{q}(\boldsymbol{y})$ respectivamente. Definimos los estimadores $\widehat{p}(\boldsymbol{x})$ y $\widehat{q}(\boldsymbol{y})$ como*

$$\widehat{p}(\boldsymbol{x}) = \sum_{i=1}^{N_{\mathbb{P}}} \pi_{s,i}^{\mathbb{P}} \widehat{b_i^{\mathbb{P}}}(\boldsymbol{x}), \quad y \quad \widehat{q}(\boldsymbol{y}) = \sum_{i=1}^{N_{\mathbb{Q}}} \pi_{s,i}^{\mathbb{Q}} \widehat{b_i^{\mathbb{Q}}}(\boldsymbol{y}), \text{ con } \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^D, \tag{3.19}$$

donde $\pi_{s,i}^{\mathbb{P}}$ y $\pi_{s,j}^{\mathbb{Q}}$ son probabilidades estacionarias de las distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$, respectivamente y los estimadores $\widehat{b_i^{\mathbb{P}}}(\boldsymbol{x})$ y $\widehat{b_i^{\mathbb{Q}}}(\boldsymbol{y})$ de $b_i^{\mathbb{P}}(\boldsymbol{x})$ y $b_i^{\mathbb{Q}}(\boldsymbol{y})$ se definen

$$\widehat{b_i^{\mathbb{P}}}(\boldsymbol{x}) = \sum_{j=1}^{M_{\mathbb{P}}} \alpha_{i,j} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}), \quad y \quad \widehat{b_i^{\mathbb{Q}}}(\boldsymbol{y}) = \sum_{j=1}^{M_{\mathbb{Q}}} \beta_{i,j} \mathcal{N}(\boldsymbol{y}|\boldsymbol{\nu}_{i,j}, \boldsymbol{\Lambda}_{i,j}).$$

Aquí, $\beta_{i,j}$ es el coeficiente para la j-ésima componente de la mezcla en el estado i , $\boldsymbol{\nu}_{i,j}$ es un parámetro de media, y $\boldsymbol{\Lambda}_{i,j}$ es el parámetro de covarianza para la componente j del estado i.

A continuación presentamos dos resultados sobre la construcción de nuevas métricas entre HMMs. Estos resultados son consecuencia de los Teoremas 3.1.1 y 3.1.5.

Corolario 3.4.2. *Si reemplazamos las expresiones de la Ecuación* (3.19) *en la Ecuación* (2.11)*, y si asumimos un kernel Gaussiano $k(\boldsymbol{x},\boldsymbol{y};\ell) = \exp\left(-\ell\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right)$, entonces obtenemos una métrica entre HMMs basada en los RKHS, la cual viene dada por*

$$\begin{aligned}\widehat{\gamma_k^2}(\mathbb{P},\mathbb{Q}) = &\sum_{i,j=1}^{N_{\mathbb{P}}} \sum_{k,l=1}^{M_{\mathbb{P}}} \pi_{s,i}^{\mathbb{P}} \pi_{s,j}^{\mathbb{P}} \alpha_{i,k}\alpha_{j,l} \widehat{k}\left(\boldsymbol{\mu}_{i,k}, \boldsymbol{\mu}_{j,l}; \boldsymbol{\Sigma}_{i,k}, \boldsymbol{\Sigma}_{j,l}, \ell\right) \\ &+ \sum_{i,j=1}^{N_{\mathbb{Q}}} \sum_{k,l=1}^{M_{\mathbb{Q}}} \pi_{s,i}^{\mathbb{Q}} \pi_{s,j}^{\mathbb{Q}} \beta_{i,k}\beta_{j,l} \widehat{k}\left(\boldsymbol{\nu}_{i,k}, \boldsymbol{\nu}_{j,l}; \boldsymbol{\Lambda}_{i,k}, \boldsymbol{\Lambda}_{j,l}, \ell\right) \\ &- 2 \sum_{i,j=1}^{N_{\mathbb{P}},N_{\mathbb{Q}}} \sum_{k,l=1}^{M_{\mathbb{P}},M_{\mathbb{Q}}} \pi_{s,i}^{\mathbb{P}} \pi_{s,j}^{\mathbb{Q}} \alpha_{i,k}\beta_{j,l} \widehat{k}\left(\boldsymbol{\mu}_{i,k}, \boldsymbol{\nu}_{j,l}; \boldsymbol{\Sigma}_{i,k}, \boldsymbol{\Lambda}_{j,l}, \ell\right),\end{aligned} \tag{3.20}$$

donde

$$\begin{aligned}\widehat{k}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\Sigma},\boldsymbol{\Lambda},\ell) &= \widehat{k}_G(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\Sigma},\boldsymbol{\Lambda},\ell) \\ &= \frac{(|\mathbf{I}\ell|)^{1/2}}{(|\boldsymbol{\Sigma}+\boldsymbol{\Lambda}+\mathbf{I}\ell|)^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{y})^\top(\boldsymbol{\Sigma}+\boldsymbol{\Lambda}+\mathbf{I}\ell)^{-1}(\boldsymbol{x}-\boldsymbol{y})}{2}\right).\end{aligned} \tag{3.21}$$

Corolario 3.4.3. *Si reemplazamos las expresiones de la Ecuación* (3.19) *en la Ecuación* (2.11) *para* $D = 1$*, y si asumimos un kernel Laplaciano* $k(x, y; \ell) = \exp(-\ell\|x - y\|_2)$*, entonces obtenemos una nueva métrica entre HMMs, la cual viene dada por*

$$\begin{aligned}\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q}) = &\sum_{i,j=1}^{N_{\mathbb{P}}} \sum_{k,l=1}^{M_{\mathbb{P}}} \pi_{s,i}^{\mathbb{P}} \pi_{s,j}^{\mathbb{P}} \alpha_{i,k} \alpha_{j,l} \widehat{k}\left(\mu_{i,k}, \mu_{j,l}; \Sigma_{i,k}, \Sigma_{j,l}, \ell\right) \\ &+ \sum_{i,j=1}^{N_{\mathbb{Q}}} \sum_{k,l=1}^{M_{\mathbb{Q}}} \pi_{s,i}^{\mathbb{Q}} \pi_{s,j}^{\mathbb{Q}} \beta_{i,k} \beta_{j,l} \widehat{k}\left(\nu_{i,k}, \nu_{j,l}; \Lambda_{i,k}, \Lambda_{j,l}, \ell\right) \\ &- 2 \sum_{i,j=1}^{N_{\mathbb{P}}, N_{\mathbb{Q}}} \sum_{k,l=1}^{M_{\mathbb{P}}, M_{\mathbb{Q}}} \pi_{s,i}^{\mathbb{P}} \pi_{s,j}^{\mathbb{Q}} \alpha_{i,k} \beta_{j,l} \widehat{k}\left(\mu_{i,k}, \nu_{j,l}; \Sigma_{i,k}, \Lambda_{j,l}, \ell\right),\end{aligned} \tag{3.22}$$

donde los parámetros $\pi_{s,i},\ \alpha_{i,k},\ \beta_{i,k},\ \mu_{i,k},\ \nu_{i,k},\ \Lambda_{i,k},\ \Sigma_{i,k} \in \mathbb{R},\ \ell\ y$

$$\begin{aligned}\widehat{k}(x, y; \Sigma, \Lambda, \ell) = &f_1(x, y; \Sigma, \Lambda, \ell) \times \left(\frac{\Sigma\sqrt{\pi}}{\Lambda} - \frac{\Sigma\sqrt{\pi}}{\Lambda} \operatorname{erf}\left(\frac{d_1(x, y; \Sigma, \Lambda, \ell)\, \Sigma}{2\Lambda\sqrt{\Lambda^2 + \Sigma^2}}\right)\right) \\ &+ f_2(x, y; \Sigma, \Lambda, \ell) \times \left(\frac{\Lambda\sqrt{\pi}}{\Sigma} - \frac{\Lambda\sqrt{\pi}}{\Sigma} \operatorname{erf}\left(\frac{d_2(x, y; \Sigma, \Lambda, \ell)\, \Lambda}{2\Sigma\sqrt{\Lambda^2 + \Sigma^2}}\right)\right),\end{aligned}$$

$$\begin{aligned}f_1(x, y; \Sigma, \Lambda, \ell) = &\Lambda^2\sqrt{\pi} \exp\left(\frac{\left(2\ell^2 x + \Lambda^2\right)^2}{8\ell^4\Lambda^2}\left(1 - \frac{\Lambda^2}{\Sigma^2}\right)\right) \\ &\times \exp\left(-\frac{\left(2\ell^2 y + \Lambda^2\right)}{2\ell^2}\left(\frac{1}{2\ell^2} - \frac{x}{\Sigma^2}\right)\right) \\ &\times \left(-\frac{1}{2}\left(\frac{y^2}{\Lambda^2} + \frac{x^2}{\Sigma^2}\right) + \frac{\Sigma^2 d_1^2(x, y; \Sigma, \Lambda, \ell)}{4\Lambda^2}\right),\end{aligned}$$

$$d_1(x, y; \Sigma, \Lambda, \ell) = \frac{\Lambda}{\sqrt{2}l^2\Sigma^2}\left(2\ell^2(x - y) + \Sigma^2 + \Lambda^2\right),$$

$$f_2(x, y; \Sigma, \Lambda, \ell) = f_1(y, x; \Lambda, \Sigma, \ell)$$

y

$$d_2(x, y; \Sigma, \Lambda, \ell) = d_1(x, y; \Lambda, \Sigma, \ell).$$

Note que

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left(-u^2\right) du,$$

es la función error de Gauss.

Nos referimos a la métrica 3.20 como $\widehat{\gamma_{k_G}^2}(\mathbb{P}, \mathbb{Q})$, y a la métrica 3.22 como $\widehat{\gamma_{k_L}^2}(\mathbb{P}, \mathbb{Q})$.

Note que las métricas entre HMMs basadas en los RKHS tienen una forma cerrada, la cual depende del hiperparámetro ℓ; del número de estados ocultos, $N_{\mathbb{P}}$ y $N_{\mathbb{Q}}$; y del número de componentes de las mezclas Gaussianas $M_{\mathbb{P}}$ y $M_{\mathbb{Q}}$. Los parámetros $\{\pi^{\mathbb{P}}_{s,i}, \pi^{\mathbb{Q}}_{s,i}, \alpha_{i,j}, \beta_{i,j}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\nu}_{j,l}; \boldsymbol{\Sigma}_{i,k}, \boldsymbol{\Lambda}_{j,l}\}$ son estimados mediante el algoritmo EM para HMMs. Este algoritmo estima los parámetros del modelo HMM iterativamente en dos pasos: la etapa E encuentra la distribución de probabilidad para las variables aleatorias no observables, dado los valores conocidos para las variables aleatorias observables y la estimación actual de los parámetros, y la etapa M reestima los parámetros del modelo, para que sean aquellos que maximizan la verosimilitud, bajo el supuesto de que la distribución de probabilidad encontrada en la etapa E es correcta [32, 31, 34].

A continuación presentamos una propiedad de las métricas entre entre HMMs estacionarios basadas en los RKH. Auque la propiedad se presenta para observaciones univariadas, sin embargo, el resultado es análogo para observaciones multivariantes.

Observación 3.4.4. *Suponga que* $\mathbb{E}_X[k_X] = \int_{\mathcal{X}} k_x p(x)dx$ *es el operador del embebimiento en el espacio de Hilbert,* $k(x,y) \leq 1$*,*

$$\widehat{p}(x) = \sum_{i=1}^{N_{\mathbb{P}}} \sum_{j=1}^{M_{\mathbb{P}}} \pi^{\mathbb{P}}_i \alpha_{i,j} \mathcal{N}(x|\mu_{i,j}, \Sigma_{i,j}),$$

un estimador de la mezcla Gaussiana de $p(x)$ *y*

$$\widehat{\mathbb{E}}_X[k_X] = \sum_{i=1}^{N_{\mathbb{P}}} g(\mu_{i,j}, \Sigma_{i,j}),$$

un estimador de $\mathbb{E}_X[k_X)]$ *donde*

$$g(\mu_{i,j}, \Sigma_{i,j}) = \sum_{j=1}^{M_{\mathbb{P}}} \pi^{\mathbb{P}}_i \alpha_{i,j} \int_{\mathcal{X}} k_x \mathcal{N}(x|\mu_{i,j}\Sigma_{i,j})dx. \tag{3.23}$$

Si $\|g(\mu_{i,j}, \Sigma_{i,j})\|_{\mathcal{H}} \leq 1$*, entonces* $\widehat{\gamma^2_k}(\mathbb{P}, \mathbb{Q})$ *es un estimador consistente de* $\gamma^2_k(\mathbb{P}, \mathbb{Q})$ *donde* $\mu_{i,j} = f(x_1, \cdots, x_{N_x})$*,* $\Sigma_{i,j} = h(x_1, \cdots, x_{N_x})$*, y* $x_1, \cdots, x_{N_x}$ *es una muestra tomada i.i.d. de una variable aleatoria* X*.*

Ciertamente, por el Teorema 7 presentado en [16], el estimador $\widehat{\gamma^2_k}(\mathbb{P}, \mathbb{Q})$ *es consistente si* $\|g(\mu_{i,j}, \Sigma_{i,j})\|_{\mathcal{H}} \leq 1$*. Verifiquemos esta última desigualdad usando la expresión*

$$\begin{aligned} \|g(\mu_{i,j}, \Sigma_{i,j})\|^2_{\mathcal{H}} &= \sum_{j=1}^{M_{\mathbb{P}}} \sum_{l=1}^{M_{\mathbb{P}}} \pi^{\mathbb{P}}_i \pi^{\mathbb{P}}_i \alpha_{i,j} \alpha_{i,l} \int_{\mathcal{X}} \int_{\mathcal{X}} k(x,y) \\ &\times \mathcal{N}(x|\mu_{i,j}, \Sigma_{i,j}) \mathcal{N}(y|\mu_{i,l}\Sigma_{i,l}) dxdy. \end{aligned} \tag{3.24}$$

Dado que $k(x,y) \leq 1$*,* $\pi^{\mathbb{P}}_i \leq 1$*,* $\sum_{j=1}^{M_{\mathbb{P}}} \alpha_{i,j} = \sum_{l=1}^{M_{\mathbb{P}}} \alpha_{i,l} = 1$*, y*

$$\int_{\mathcal{X}} \mathcal{N}(x|\mu_{i,j}, \Sigma_{i,j})dx = \int_{\mathcal{X}} \mathcal{N}(y|\mu_{i,l}, \Sigma_{i,l})dy = 1,$$

entonces $\|g(\mu_{i,j}, \Sigma_{i,j})\|_{\mathcal{H}} \leq 1$*.*

En esta sección, comparamos las métricas entre HMMs que proponemos en este trabajo con la medida KL. Para el análisis de los resultados en clasificación de series de tiempo, usamos las bases de datos binarias y el concepto de coeficiente de variación CV multivariado definido en la Sección 3.5.2. La medida KL la definimos como en [56].

Definición 3.4.5. *Sean ρ_1 y ρ_2 dos HMMs y $X_1, X_2, \cdots, X_T$ una secuencia de longitud T generada por ρ_1, entonces la medida KL viene dada por*

$$d(\rho_1, \rho_2) = \frac{1}{T}\left(\sum_{i=1}^{T} \log p(x_i) - \log q(x_i)\right),$$

donde $p(X_i = x_i) = p(x_i)$ y $q(X_i = x_i) = q(x_i)$ son distribuciones de probabilidad de ρ_1 y ρ_2 respectivamente, y son estimados como en (3.19) donde $p(x_i) \geq q(x_i)$ dado que la secuencia es generada por ρ_1. Además, esta medida no es una métrica dado que no satisface la desigualdad triangular y no es simétrica. Sin embargo, podemos definir la versión simétrica KL de esta medida de la siguiente manera:

$$KL(\rho_1, \rho_2) = \frac{1}{2}\left(d(\rho_1, \rho_2) + d(\rho_2, \rho_1)\right). \tag{3.25}$$

3.5 Resultados experimentales de las métricas entre HMMs basadas en RKHS

En esta sección, describimos brevemente la evaluación experimental empleada cuando las métricas basadas en los RKHS que proponemos en este trabajo, son utilizadas para discriminar dos HMMs usando datos sintéticos y datos disponibles en el archivo clasificación UCR de series de tiempo. Evaluamos el rendimiento de las métricas en ambos casos utilizando el clasificador KNN, con tres vecinos ($K = 3$), para los datos sintéticos y con uno, tres y cinco vecinos ($K = 1, 3, 5$) para las bases de datos de la UCR.

3.5.1 Datos sintéticos

Con el fin evaluar el rendimiento de las métricas propuestas en este trabajo, comparamos dos HMMs sintéticos con distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$ cada uno. Fijamos los mismos parámetros para ambos HMM, pero cambiamos la matriz de transición. Evaluamos el rendimiento de las métricas basadas en RKHS para series de tiempo de longitud $T = 200, 500$ y para el estado oculto ($N_{\mathbb{P}} = N_{\mathbb{Q}} = 3$). Para las probabilidades de emisión, asumimos un modelo de mezclas de Gaussianas con una sola componente Gaussiana, es decir, $M_{\mathbb{P}} = M_{\mathbb{Q}} = 1$. En este experimento, diseñamos un modelo HMM base para la distribución $\mathbb{P}$, fijamos una matriz de transición $A^{\mathbb{P}}$. Para la distribución $\mathbb{Q}$, fijamos el mismo parámetro empleado para la distribución $\mathbb{P}$ pero la matriz de transición $A^{\mathbb{Q}}$ es cambiada poco a poco. Luego, fijamos los valores de los parámetros media y covarianza, y generamos 200 muestras del HMM obtenido. Después de generar las muestras, empleamos el algoritmo EM [31, 34] para la estimación de los parámetros del HMM cuando se cambia la longitud de la serie de tiempo. Repetimos el mismo procedimiento para la estimación de los parámetros del HMM para la distribución $\mathbb{Q}$ que el empleado para la distribución $\mathbb{P}$.

Definimos $\boldsymbol{\mu}^{\mathbb{P}} = \boldsymbol{\mu}^{\mathbb{Q}} = [1, \quad 3, \quad 5]$, $\boldsymbol{\Sigma}^{\mathbb{P}} = \boldsymbol{\Sigma}^{\mathbb{Q}} = \text{diag}(0.2)$, y la matriz de transición base $A^{\mathbb{P}}$ dada por

$$A^{\mathbb{P}} = \begin{bmatrix} 0.60 & 0.20 & 0.20 \\ 0.20 & 0.60 & 0.20 \\ 0.20 & 0.20 & 0.60 \end{bmatrix}.$$

Variaciones de la matriz de transición $A^{\mathbb{Q}}$

$$A_1^{\mathbb{Q}} = \begin{bmatrix} 0.46 & 0.27 & 0.27 \\ 0.27 & 0.46 & 0.27 \\ 0.27 & 0.27 & 0.46 \end{bmatrix}, \; A_2^{\mathbb{Q}} = \begin{bmatrix} 0.20 & 0.40 & 0.40 \\ 0.40 & 0.20 & 0.40 \\ 0.40 & 0.40 & 0.20 \end{bmatrix},$$
$$A_3^{\mathbb{Q}} = \begin{bmatrix} 0.36 & 0.16 & 0.48 \\ 0.16 & 0.68 & 0.16 \\ 0.48 & 0.16 & 0.36 \end{bmatrix}, \; A_4^{\mathbb{Q}} = \begin{bmatrix} 0.20 & 0.20 & 0.60 \\ 0.20 & 0.60 & 0.20 \\ 0.60 & 0.20 & 0.20 \end{bmatrix}.$$

Usamos 70 muestras (elegidas al azar) por cada clase para entrenar al clasificador KNN con $K = 3$. Las 30 muestras restantes por clase son utilizadas para probar el rendimiento de las métricas. Los valores de ℓ para los kernels Gaussiano y Laplaciano se estiman mediante validación cruzada usando los datos de entrenamiento. Elegimos al azar el 40% de los datos de entrenamiento para entrenar al clasificador KNN, y evaluamos su desempeño usando el otro 60% de los datos de entrenamiento para elegir el mejor valor de ℓ que presenta los resultados con mayor precisión en el paso de la clasificación. El rendimiento de precisión se calcula mediante la suma de los casos exitosos, y dividiendo el resultado por el número total de muestras de prueba. Este procedimiento se repite diez veces, y se calculan la media y la desviación estándar ($\mu \pm \sigma$).

A continuación presentamos los resultados de clasificación de HMMs para las métricas basadas en embebimiento con kernel Gaussiano (KEG) y embebimiento con kernel Laplaciano (KEL) y la medida KL.

Tabla 3.4: Resultados de precisión usando el algoritmo KNN para las métricas KEG y KEL y la medida KL con longitud de secuencia $T_{\mathbb{P}} = T_{\mathbb{Q}} = 200$ con $Q = 3$ y $K = 3$. La media μ y la desviación estándar σ se muestran para diez repeticiones de cada experimento ($\mu \pm \sigma$).

	KL	KEG	KEL
$A_1^{\mathbb{Q}}$	0.7200 ± 0.0470	0.7183 ± 0.0487	$\mathbf{0.7300 \pm 0.0537}$
$A_2^{\mathbb{Q}}$	0.5833 ± 0.0324	0.5883 ± 0.0516	$\mathbf{0.6050 \pm 0.0423}$
$A_3^{\mathbb{Q}}$	$\mathbf{0.6983 \pm 0.0606}$	0.6167 ± 0.0633	0.6950 ± 0.0458
$A_4^{\mathbb{Q}}$	0.6883 ± 0.0478	0.6317 ± 0.0547	$\mathbf{0.6883 \pm 0.0445}$

Las Tablas 3.4 y 3.5 muestran los resultados obtenidos. Observamos que las métricas KEL y KEG presentan mejor rendimiento en clasificación que la medida KL. En general, observamos que la métrica KEL tiene un mejor rendimiento que la métrica KEG, esto se debe a que el decaimiento lento del kernel Laplaciano, captura mejor los valores cercanos a cero [14]. Notamos también que para el HMM con matriz de transición $A_2^{\mathbb{Q}}$ con respeto a la métrica KEL, obtenemos una precisión de clasificación diferente para los casos ($T = 200; 500$). Sin embargo, para el modelo con matriz de transición $A_4^{\mathbb{Q}}$

Tabla 3.5: La descripción de la tabla es la misma que la Tabla 3.4. Aquí, la longitud de secuencia $T_{\mathbb{P}} = T_{\mathbb{Q}} = 500$ se usa con $Q = 3$ y $K = 3$

	KL	KEG	KEL
$A_1^{\mathbb{Q}}$	0.5950 ± 0.0578	0.5733 ± 0.0370	$\mathbf{0.6417 \pm 0.0517}$
$A_2^{\mathbb{Q}}$	0.5750 ± 0.0379	0.5667 ± 0.0430	$\mathbf{0.6317 \pm 0.0372}$
$A_3^{\mathbb{Q}}$	0.6200 ± 0.0443	0.6167 ± 0.0633	$\mathbf{0.6583 \pm 0.0371}$
$A_4^{\mathbb{Q}}$	$\mathbf{0.6367 \pm 0.0520}$	0.6317 ± 0.0547	0.6167 ± 0.0430

sucede lo contrario, esto es porque la matriz $A^{\mathbb{P}}$ está más cerca de $A_4^{\mathbb{Q}}$. Para modelos con matrices $A_1^{\mathbb{Q}}$ y $A_3^{\mathbb{Q}}$ la métrica KEL tiene diferente precisión en clasificación cuando $T = 500$, para el caso $T = 200$ la métrica KEL tiene la misma precisión en clasificación. Estos análisis muestran que en la mayoría de los casos, los resultados obtenidos con la métrica KEG no son diferentes de los obtenidos con la medida KL. Para la métrica KEL, los resultados son generalmente diferentes de las otras medidas.

3.5.2 Base de datos de la UCR

En esta sección, se evalúa el desempeño de las métricas basadas en el método de embebimiento distribuciones de probabilidad en un RKHS usando diferentes conjuntos de datos binarios del archivo clasificación UCR de series de tiempo. En la Tabla 3.1 se describe las bases de datos de la UCR que utilizamos en esta sección. El CV que definimos en 3.2.1 es un concepto que usamos en el análisis de los resultados de clasificación de series de tiempo para las bases de datos binarias.

En los experimentos, se supone que $M_{\mathbb{P}} = M_{\mathbb{Q}} = 1$ y para la selección del número de estados ocultos para cada una de las bases de datos, implementamos el método de selección estratégica Criterio de Inferencia Bayesiana (BIC). La idea fundamental del método consiste en maximizar la probabilidad de los datos mientras se penalizan los modelos de gran tamaño, es decir, reducir el tamaño de los parámetros del modelo. En el método BIC, el número óptimo de estados N es el maximizador de la función $BIC(N) = \log(\mathbf{x}|\widehat{\rho}_N) - (R_N/2)\log(T)$ donde, $\mathbf{x}$ es el conjunto de datos observados, T es el número total de observaciones en $\mathbf{x}$, $\widehat{\rho}_N$ indica la estimación de máxima verosimilitud del modelo HMM con N estados, y R_N es la cantidad total de parámetros libres de $\widehat{\rho}_N$ [3].

La Tabla 3.6 muestra el rendimiento de las métricas basadas en los RKHS sobre los conjuntos de datos disponibles en la UCR. De acuerdo con la Tabla 3.6, la métrica KEL tiene un mejor rendimiento que la métrica KEG y la medida KL en el 58.06%, 51.61% y 51.61% de las bases de datos UCR, usando el clasificador KNN para $K = 1, 3, 5$ respectivamente. También observamos que la métrica KEL tiene mejor rendimiento que la medida KL en el 70.97%, 61.29% y 58.06% de las bases de datos de la UCR para $K = 1, 3, 5$ respectivamente. De la Tabla 3.6 se puede concluir que la métrica KEG tiene mejor rendimiento que la medida KL en el 54.84%, 54.84% y 51.61% de las bases de datos UCR para $K = 1, 3, 5$ respectivamente. Finalmente, la métrica KEL tiene un mejor rendimiento que la métrica KEG en el 73.33%, 63.33% y 64.29% de las bases de

Tabla 3.6: Comparación del rendimiento de las métricas KEL y KEG con respecto a la medida KL usando el algoritmo KNN para $K = 1, 3, 5$

	K=1			K=3			K=5		
	KL	KEG	KEL	KL	KEG	KEL	KL	KEG	KEL
BeetleFly	0.8500	0.7000	0.6500	0.8500	0.8500	0.5500	0.8500	0.8000	0.8000
BirdChicken	0.7500	0.8500	0.8500	0.8000	0.8500	0.8500	0.7500	0.7500	0.9000
Coffee	0.6429	0.5714	0.7500	0.5714	0.6786	0.7857	0.6429	0.7500	0.6786
Computers	0.5240	0.6280	0.6760	0.5920	0.6480	0.6240	0.6120	0.6600	0.6480
DistalPhalanxOutlineCorrect	0.5960	0.6267	0.6367	0.5750	0.6717	0.6783	0.5417	0.6817	0.6817
ECG200	0.4900	0.5500	0.6600	0.5700	0.6300	0.6700	0.6300	0.6800	0.7200
ECGFiveDays	0.6016	0.7584	0.7758	0.7573	0.7735	0.7944	0.7944	0.7898	0.8014
Earthquakes	0.6957	0.7019	0.8199	0.7764	0.7671	0.8199	0.7888	0.7733	0.8199
FordA	0.5110	0.5082	0.5254	0.5369	0.5282	0.5321	0.5412	0.5121	0.5190
FordB	0.4816	0.5226	0.5154	0.5300	0.5415	0.5234	0.5430	0.5333	0.5105
GunPoint	0.9133	0.7467	0.7933	0.8667	0.6933	0.6800	0.7867	0.6667	0.6667
Ham	0.4857	0.4857	0.5238	0.4857	0.5714	0.4952	0.4381	0.6190	0.5238
HandOutlines	0.6110	0.5910	0.6120	0.6510	0.6340	0.6470	0.6610	0.6620	0.6690
Herring	0.5156	0.4062	0.5938	0.5156	0.4062	0.5938	0.5938	0.4844	0.5938
ItalyPowerDemand	0.5063	0.7123	0.6929	0.5500	0.7143	0.7454	0.5355	0.7211	0.7259
Lighting2	0.7377	0.7049	0.7213	0.7377	0.7213	0.6721	0.7377	0.7213	0.6721
MiddlePhalanxOutlineCorrect	0.3750	0.5783	0.6133	0.3850	0.5483	0.6450	0.4617	0.5617	0.6450
MoteStrain	0.7524	0.7276	0.7324	0.7556	0.7300	0.7212	0.7764	0.6717	0.5391
PhalangesOutlinesCorrect	0.4009	0.5874	0.6585	0.3939	0.6166	0.6865	0.3963	0.6247	0.6911
ProximalPhalanxOutlineCorrect	0.3058	0.7079	0.7113	0.3058	0.7320	0.7766	0.3196	0.7457	0.7629
ShapeletSim	0.5111	0.4500	0.5000	0.4389	0.4833	0.5000	0.4500	0.4778	0.5000
SonyAIBORobotSurface	0.8136	0.6639	0.4925	0.7970	0.6872	0.6173	0.7687	0.7155	0.4792
SonyAIBORobotSurfaceII	0.7230	0.5603	0.6128	0.7303	0.6243	0.5992	0.7125	0.6170	0.5813
Strawberry	0.4845	0.6933	0.7635	0.4878	0.7096	0.7439	0.4878	0.7145	0.7243
ToeSegmentation1	0.7061	0.6974	0.6535	0.7719	0.7149	0.6184	0.7675	0.6930	0.7149
ToeSegmentation2	0.5077	0.6077	0.6154	0.6000	0.6846	0.7077	0.6077	0.7154	0.7662
TwoLeadECG	0.5909	0.7032	0.6883	0.7191	0.6901	0.6962	0.7937	0.6304	0.7270
Wine	0.5185	0.6111	0.5926	0.4815	0.5741	0.6296	0.5370	0.4815	0.6296
WormsTwoClass	0.5138	0.5801	0.5856	0.5083	0.5912	0.5691	0.4972	0.5801	0.5746
Wafer	0.7583	0.9283	0.9517	0.9113	0.9356	0.9445	0.9189	0.9332	0.9395
Yoga	0.6780	0.6770	0.6653	0.7427	0.7410	0.6680	0.7407	0.7493	0.6653
Ganador	9/31	5/31	**18/31**	11/31	6/31	**16/31**	11/31	5/31	**16/31**

datos UCR, para $K = 1, 3, 5$ respectivamente.

De acuerdo a la Tabla 3.6, también observamos que las métricas basadas en el método de embebimiento de distribuciones de probabilidad en un RKHS KEL y KEG tienen mejor rendimiento en clasificación que la medida KL (KEL,KEG vs KL) en el 70%, 66.67% y 66.67% de las bases de datos UCR, usando el clasificador KNN para $K = 1, 3, 5$ (respectivamente). Estos resultados también muestran que la métrica KEL tiene su mejor rendimiento cuando $K = 1$, cuando $K = 3, 5$ el rendimiento de la métrica KEL disminuye ligeramente. El rendimiento de las medidas KL y KEG no cambia mucho para $K = 1, 3, 5$.

Los resultados presentados en la Tabla 3.6 muestran que las bases de datos donde las métricas KEL y KEG tienen mejor rendimiento que la medida KL incluye las bases de datos de la Tabla 3.7. En estas bases de datos se puede observar que cuando el CV de los datos de entrenamiento y prueba son pequeños, al menos una de las métricas KEL o KEG tienen un rendimiento notablemente mejor que la medida KL. El CV para estas bases de datos se representa en la Tabla 3.7. La razón fundamental de este rendimiento competitivo de las métricas KEL y KEG es porque ellas se basan en kernels suaves. Cuando estas métricas se entrenan con un conjunto de series de tiempo que tienen un CV pequeño, es decir series temporales con poca variabilidad, entonces las métricas KEL y KEG tienen un mejor rendimiento en clasificación cuando los conjuntos

Tabla 3.7: Bases de datos donde el CV de los conjuntos de entrenamiento y de prueba son pequeños

Base de datos	CV datos de entrenamiento	CV de datos de prueba
Coffee	0.0955	0.1052
DistalPhalanxOutlineCorrect	0.2820	0.2722
ECG200	0.7542	0.9026
ItalyPowerDemand	0.5911	0.5303
MiddlePhalanxOutlineCorrect	0.1533	0.1620
ProximalPhalanxOutlineCorrect	0.2074	0.2117
TwoLeadECG	0.4240	0.4186
Wine	0.2000	0.0252

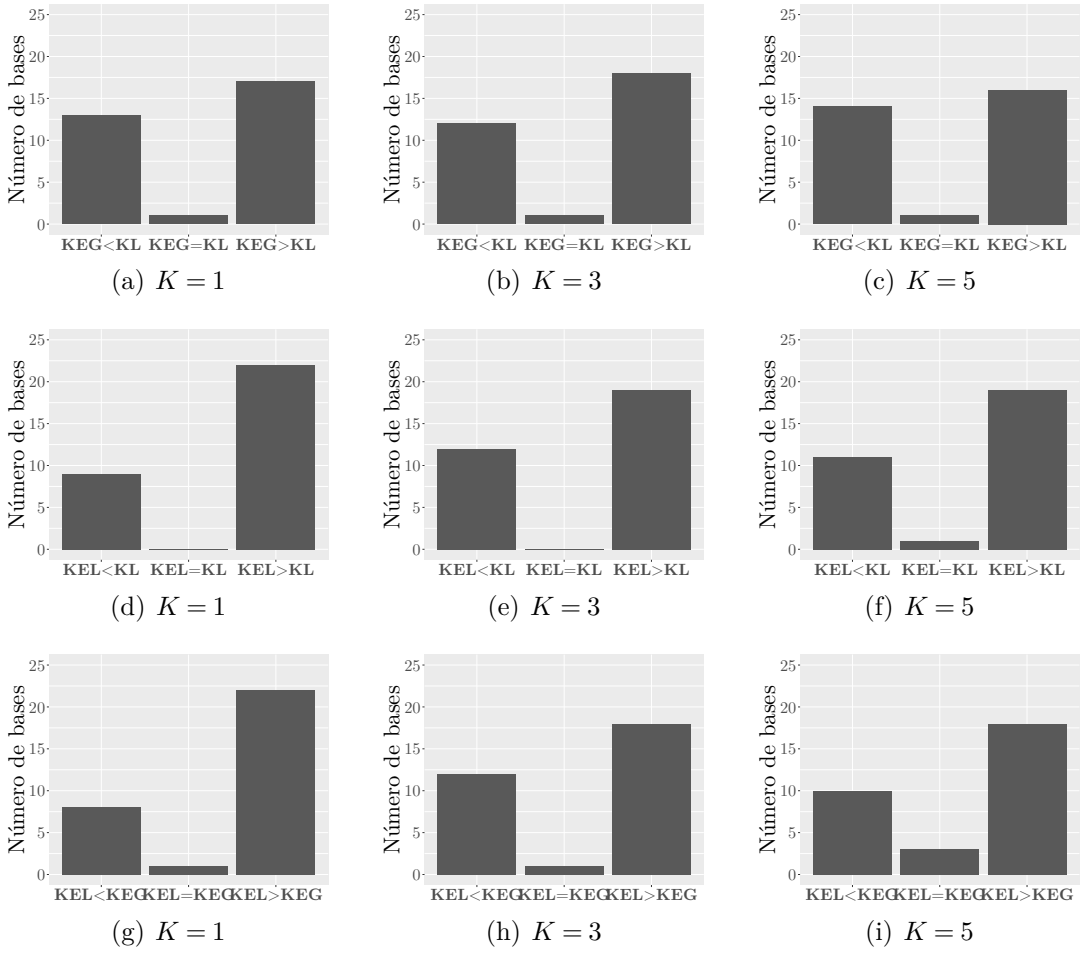
de prueba tienen poca variabilidad.

De acuerdo a la Tabla 3.6, se observa que cuando al menos uno de los CV de los conjuntos de entrenamiento y de prueba es mayor que uno, es decir las series de tiempo de estas bases de datos tienen alta variabilidad, entonces las métricas KEL y KEG muestran un rendimiento menor que la medida KL. Por ejemplo, para la base de datos BeetleFly el CV de los conjuntos de entrenamiento y prueba son 2.8510 y 2.1903 (respectivamente), para la base de datos Lighting son 1.5305 y 1.5318 (respectivamente), mientras que para ShapeletSim MoteStrain son 1.1358 y 1.0404 (respectivamente), SonyAIBORobotSurface es 13.5086 y 4.9382 (respectivamente), para ToeSegmentation1 son 6.0772 y 1.4092 (respectivamente) y para Yoga son 1.7160 y 1.6958 (respectivamente). La razón puede ser porque las méricas KEL y KEG se basan en kernels suaves, y cuando estas métricas son entrenadas con series de tiempo que tienen alta variabilidad, entonces el rendimiento en clasificación disminuye. Esta puede ser la razón principal por la cual la medida KL tiene un mejor rendimiento que las métricas KEL y KEG para estas bases de datos.

También observamos de la Tabla 3.6, que las medidas KEL, KEG y KL tienen bajo rendimiento en clasificación de series de tiempo, cuando el CV de los conjuntos de entrenamiento y prueba son muy altos. Por ejemplo, para la base de datos FordA el CV de los conjuntos de entrenamiento y prueba son 11.5505 y 16.1110 (respectivamente), para FordB son 29.8144 y 57.8149, mientras que para la base de datos ShapeletSim son 8.6221 y 1.0404 y para la base de datos WormsTwoClass son 4.9382 y 13.3578 . La razón puede ser porque las tres medidas se basan en HMM estacionarios, y esta hipótesis está lejos de cumplirse en estas bases de datos con un CV muy alto.

Finalmente, se observó de la Tabla 3.6, que la métrica KEL tiene un mejor rendimiento que la métrica KEG, esto se debe a que la secuencia de valores propios del kernel Laplaciano converge más rápido a cero, lo cual implica que este kernel mejora la detección de la dependencia codificada de las frecuencias más altas en la distribución de probabilidad.

Las Figuras 3.2(a), (d) y (g) muestran el número de bases de datos donde la métrica KEL tiene mejor rendimiento que la medida KL, donde ellas son iguales y donde la medida KL tiene mejor rendimiento que la métrica KEL para $K = 1, 3, 5$ respectivamente. Análogamente, las Figuras 3.2(b), (e) y (h) muestran el número de bases de datos



(a) $K = 1$ (b) $K = 3$ (c) $K = 5$

(d) $K = 1$ (e) $K = 3$ (f) $K = 5$

(g) $K = 1$ (h) $K = 3$ (i) $K = 5$

Figura 3.2: La figura muestra el rendimiento en clasificación de series de tiempo de las métricas KEL y KEG en comparación con la medida KL utilizando el algoritmo KNN pora $K = 1,3,5$ para las treinta y una bases de datos binarias del Archivo de Clasificación de Series Temporales UCR.

donde la métrica KEL es mayor, igual y menor que la métrica KEG para $K = 1,3,5$. Finalmente, las Figuras 3.2(c), (f) y (g) muestran el número de bases de datos donde la métrica KEG es mayor, igual y menor que la medida KL para $K = 1,3,5$.

Finalmente, en esta sección comparamos y analizamos el rendimiento de la métrica PKE entre distribuciones de probabilidad basada en un RKHS y las métricas KEG y KEL entre HMMs basadas en un RKHS. Todas las métricas son evaluadas en 24 bases de datos binarias de la UCR, utilizando el algoritmo KNN, para $K = 1,3,5$.

En general, podemos decir que los resultados obtenidos con las métricas entre HMMs basadas en embebimiento de distribuciones de probabilidad en un RKHS muestran un buen rendimiento en clasificación de series de tiempo usando el algoritmo KNN para $K = 1,3,5$, si se comparan con la medida KL. Sin embargo, al comparar el rendimiento de las métricas KEL y KEG con la métrica entre distribuciones de probabilidad basada en los RKHS y el estimador de densidad de Parzen PKE, observamos que las métricas KEL y KEG tienen un menor rendimiento en clasificación de series de tiempo que

Tabla 3.8: Comparación del rendimiento de las métricas KEL y KEG con respecto a la medida PKE usando el algoritmo KNN para $K = 1, 3, 5$

	K=1			K=3			K=5		
	PKE	KEG	KEL	PKE	KEG	KEL	PKE	KEG	KEL
BeetleFly	0.7500	0.7000	0.6500	0.7000	0.8500	0.5500	0.7500	0.8000	0.8000
BirdChicken	0.9500	0.8500	0.8500	0.8500	0.8500	0.8500	0.8500	0.7500	0.9000
Coffee	0.8214	0.5714	0.7500	0.8214	0.6786	0.7857	0.8571	0.7500	0.6786
DistalPhalanxOutlineCorrect	0.7417	0.6267	0.6367	0.7450	0.6717	0.6783	0.7750	0.6817	0.6817
ECG200	0.7800	0.5500	0.6600	0.7700	0.6300	0.6700	0.7800	0.6800	0.7200
ECGFiveDays	0.8641	0.7584	0.7758	0.8551	0.7735	0.7944	0.7851	0.7898	0.8014
Earthquakes	0.7142	0.7019	0.8199	0.7671	0.7671	0.8199	0.8106	0.7733	0.8199
FordA	0.5460	0.5082	0.5254	0.5590	0.5282	0.5321	0.5740	0.5121	0.5190
FordB	0.5294	0.5226	0.5154	0.5490	0.5415	0.5234	0.5503	0.5333	0.5105
GunPoint	0.8733	0.7467	0.7933	0.8467	0.6933	0.6800	0.7533	0.6667	0.6667
Ham	0.4857	0.4857	0.5238	0.5619	0.5714	0.4952	0.6476	0.6190	0.5238
Herring	0.6250	0.4062	0.5938	0.6094	0.4062	0.5938	0.5625	0.4844	0.5938
Lighting2	0.7377	0.7049	0.7213	0.7213	0.7213	0.6721	0.7213	0.7213	0.6721
MiddlePhalanxOutlineCorrect	0.6317	0.5783	0.6133	0.6617	0.5483	0.6450	0.6583	0.5617	0.6450
ProximalPhalanxOutlineCorrect	0.7595	0.7079	0.7113	0.8076	0.7320	0.7766	0.8007	0.7457	0.7629
ShapeletSim	0.5222	0.4500	0.5000	0.5056	0.4833	0.5000	0.5056	0.4778	0.5000
SonyAIBORobotSurface	0.7820	0.6639	0.4925	0.8236	0.6872	0.6173	0.8053	0.7155	0.4792
SonyAIBORobotSurfaceII	0.7618	0.5603	0.6128	0.7786	0.6243	0.5992	0.7775	0.6170	0.5813
Strawberry	0.8825	0.6933	0.7635	0.8842	0.7096	0.7439	0.8564	0.7145	0.7243
ToeSegmentation1	0.7105	0.6974	0.6535	0.7412	0.7149	0.6184	0.7193	0.6930	0.7149
ToeSegmentation2	0.8000	0.6077	0.6154	0.8000	0.6846	0.7077	0.7385	0.7154	0.7662
TwoLeadECG	0.8973	0.7032	0.6883	0.8850	0.6901	0.6962	0.8797	0.6304	0.7270
Wine	0.7593	0.6111	0.5926	0.7222	0.5741	0.6296	0.6852	0.4815	0.6296
WormsTwoClass	0.5801	0.5801	0.5856	0.5856	0.5912	0.5691	0.5414	0.5801	0.5746
Ganador	**21/24**	0/24	3/24	**20/24**	5/24	2/24	**17/24**	3/24	6/24

la métrica PKE sobre 24 bases de datos de la tabla 3.1. En la Tabla 3.8 se puede observar que la métrica PKE tiene un mejor rendimiento que las métricas KEL y KEG en el 87.5%, 83.35% y 70.83% de las 24 bases de de datos de la UCR para $K = 1, 3, 5$ respectivamente. También podemos observar de la Tabla 3.8 que el rendimiento de las métricas (KEL, KEG) aumenta cuando el número de vecinos K crece.

Los resultados obtenidos sobre el rendimiento en clasificación de series de tiempo entre las métricas KEG, KEL y PKE son un poco sorprendentes. Antes de conocer los resultados experimentales, pensaríamos que las métricas KEG y KEL tendrían un mejor rendimiento que la métrica PKE, dado que las métricas KEG y KEL tienen en cuenta la variable temporal, mientras que la métrica PKE no la tiene encuenta. Creemos que existen tres razones que justifican estos resultados. La primera razón se debe a la suposición que hemos asumido de que los HMMs son estacionarios, es decir, la matriz de transición no depende del tiempo. La segunda razón, se debe a que las métricas entre HMMs basadas en RKHS tienen muchos parámetros a estimar, sobre todo el número de estados de los HMMs. La estimación del número de estados de los HMMs se realizo mediante el método de optimización BIC descrito en la sección 3.5.2. Este método tiene el inconveniente de generar un estimador para el número de estados donde la consistencia es un problema que aun no ha sido resuelto [38]. Esto puede implicar, que el rendimiento en clasificación de las métricas entre HMMs es mejor cuando el número de datos de entrenamiento de las bases de datos es pequeño, como por ejemplo las bases de datos Beetlefly, BirdChicken, ECGFiveDays, Herring, ToeSegmentation2 y WormsTwoClass. Para estas bases de datos las métricas (KEL, KEG) superan a la métrica PKE. La tercera razón es porque la métrica PKE tiene pocos parámetros a estimar y esta métrica depende directamente más de los datos que las métricas KEG y KEL.

3.6 Resumen y comentarios del Capítulo 3

En este capítulo, desarrollamos nuevas métricas entre distribuciones de probabilidad y entre HMMs estacionarios usando el método de embebimiento distribuciones de probabilidad en un RKHS. Estas métricas son desarrolladas de forma cerrada y se basan en la función de distribución acumulativa. Evaluamos el desempeño de las métricas basadas en los RKHS usando el algoritmo de clasificación KNN para $K = 1,\ 3,\ 5$. Los resultados experimentales presentados en este capítulo mostraron que las métricas entre distribuciones de probabilidad y entre HMMs basadas en los RKHS tienen buen rendimiento en clasificación de series de tiempo.

Una aplicación importante de los embebimientos de distribuciones de probabilidad en un RKHS es desarrollar medidas de dependencia entre variables aleatorias, las cuales han sido ampliamente utilizadas en el análisis de datos [44]. Estas medidas se pueden ver como un caso particular de una medida de distancia entre distribuciones de probabilidad en un RKHS. Aunque existen diferentes aplicaciones de medidas de dependencia entre variables aleatorias, en esta tesis nos interesa utilizar estas medidas en un modelo donde exista una relación de dependencia entre las variables aleatorias que intervienen en el modelo. En el próximo capítulo, vamos a resolver un modelo autorregresivo usando como medida de dependencia la norma del operador de covarianza cruzada.

Capítulo 4

Predicción a corto plazo de series de tiempo basada en el método embebimiento en espacios de Hilbert de procesos autorregresivos

Los procesos autorregresivos (AR) de orden p se usan habitualmente en el modelamiento y predicción de procesos aleatorios lineales estacionarios y son una herramienta fundamental en el análisis de series de tiempo que tienen aplicaciones en el pronostico de la demanda de la energía [28], en el pronostico demográfico [25] y en el pronostico de series financieras [51]. Un proceso AR de orden p, es aleatorio donde la variable de salida depende linealmente de sus propios p valores anteriores más un ruido Gaussiano [42].

Distintos autores han propuesto diversos métodos para hacer predición del proceso autorregresivo, por ejemplo el método de mínimos cuadrados, el método de máxima verosimilitud y el método de Yule-Walker. Otros autores han propuesto el uso de métodos de regresión no lineal para extender el proceso autorregresivo clásico. Estos métodos de regresión incluyen las redes neuronales [33], procesos Gaussianos [22] y métodos basados en kernel [23].

Dentro de la literatura de los procesos autorregresivos, distintos autores han extendido el marco conceptual de la investigación de estos procesos a los métodos tradicionales basados en kernel [24, 23]. En este contexto, los procesos autorregresivos basados en los métodos kernel tradicionales han sido usados para hacer pronóstico. Estos métodos kernel consisten en transformar mediante un operador no lineal, el modelo definido en un espacio de entrada a un espacio de Hilbert $\mathcal{H}$ generado por una función kernel. Lo más importante de estos modelos es que no necesitan definir explícitamente la transformación no lineal y la función kernel puede ser expresada como el producto escalar de la transformación no lineal en el espacio de Hilbert $\mathcal{H}$ [23].

Por otra parte, en [24] los autores proponen un proceso autorregresivo construido sobre un espacio de Hilbert o espacio de características. Los coeficientes del modelo autorregresivo son estimados minimizando el error de predicción entre las muestras

reales y el modelo teórico en el espacio de Hilbert. Las predicciones se calculan sólo para espacios de Hilbert de dimensión finita, con el propósito de que la función inversa definida desde el espacio de Hilbert al espacio de entrada se pueda calcular fácilmente. En [23], los autores también proponen un proceso AR construido sobre un espacio de características. En este trabajo, los coeficientes del modelo autorregresivo son estimados utilizando las ecuaciones de Yule-Walker defindas en el espacio de Hilbert. Las predicciones se obtienen resolviendo un problema de pre-imagen [18].

En este capítulo se presenta una aplicación en DSP, usando el método embebimiento de distribuciones de probabilidad en un RKHS, donde se supondrá relación de dependencia entre las observaciones, esta aplicación es una versión kernelizada de un proceso autorregresivo por medio del algoritmo de Yule-Walker, y en lugar de calcular las correlaciones (como el clásico modelo lineal) o productos punto (como en [23]), se calculan operadores de covarianza cruzada para pares de variables aleatorias. Para la predicción de series de tiempo, es necesario resolver un problema de pre-imagen [18], para mapear desde el espacio de los operadores de covarianza al espacio de entrada original. El rendimiento del modelo propuesto se compara con el modelo AR lineal, el método kernel propuesto en [23], las redes neuronales y los procesos Gaussianos. Este capítulo es basado en la publicación (*iii*)

4.1 Modelos autorregresivos en un TP-RKHS

En esta sección, se describe como el básico modelo autorregreisvo puede ser embebido en un TP-RKHS. A continuación se ofrece un método de estimación de los parámetros en el espacio de Hilbert embebido, por medio de las ecuaciones de Yule-Walker.

Definición 4.1.1. *Un proceso autorregresivo (AR) de orden p, es un proceso aleatorio donde la variable de salida depende linealmente de sus propios p valores anteriores más un ruido Gaussiano [42]. Es decir, sea $X_1, X_2, \cdots, X_n$ un proceso aleatorio discreto. Un modelo AR lineal de orden p (LAR) es definido por*

$$X_i = \lambda_1 X_{i-1} + \lambda_2 X_{i-2} + \cdots + \lambda_p X_{i-p} + \epsilon_i = \sum_{j=1}^{p} \lambda_j X_{i-j} + \epsilon_i, \tag{4.1}$$

para $i = p+1, p+2, \cdots, n$, donde $\lambda_1, \lambda_2, \cdots, \lambda_p$ son los parámetros del modelo, $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_p]^\top$ y ϵ_i es un ruido blanco con $\mathbb{E}(\epsilon_i) = 0$ y $\mathrm{var}(\epsilon_i) = \sigma^2$. Además decimos que el modelo AR es estacionario si $\mathbb{E}(\epsilon_i) = \mu$ y $\mathrm{var}(\epsilon_i) = \sigma^2$ para $i = p+1, p+2, \cdots, n$ donde μ y σ son constantes.

Las ecuaciones de Yule-Walker son un conjunto de ecuaciones lineales usadas para estimar los coeficientes $\boldsymbol{\lambda}$. La idea básica es definir un conjunto de p ecuaciones lineales, apartir de la Definición 4.1.1. Cada ecuación lineal en el sistema de Yule-Walker esta definida de la siguiente manera: sean X_i, y X_{i-k} variables aleatorias, se define la covarianza entre las dos variables aleatorias como

$$\langle X_i, X_{i-k} \rangle = \sum_{j=1}^{p} \lambda_j \langle X_{i-j}, X_{i-k} \rangle + \langle \epsilon_i, X_{i-k} \rangle, \quad \text{para } k = 1, \ldots, p. \tag{4.2}$$

Si suponemos la independencia entre ϵ_i, y X_{i-k}, el conjunto de ecuaciones lineales 4.2 se reduce a

$$\langle X_i, X_{i-k}\rangle = \sum_{j=1}^{p} \lambda_j \langle X_{i-j}, X_{i-k}\rangle, \quad \text{para } k = 1, \ldots, p. \tag{4.3}$$

Observación 4.1.2. *Dado un conjunto de observaciones para el proceso aleatorio de tiempo discreto y un estimador apropiado para los términos de covarianza $\langle X_i, X_{i-k}\rangle$, es posible solucionar el conjunto de ecuaciones para estimar $\boldsymbol{\lambda}$.*

Antes de embeber el modelo AR definido en 4.1.1 en un TP-RKHS, presentamos el procedimiento desarrollado por los autores en [23] para realizar una extensión no lineal del modelo AR utilizando un kernel en un espacio de Hilbert.

Modelo autorregresivo en un espacio de Hilbert. Este método consiste en definir las ecuaciones de Yule-Walker en el espacio de Hilbert y mostrar que los parámetros del modelo pueden ser utilizandos usando el concepto de esperanza de kernels. Supongamos que se tiene el modelo AR definido en 4.1.1, aplicando la tranformación no-lineal $\varphi : \mathcal{X} \to \mathcal{H}$ a las variables aleatorias X_i en el modelo AR,

$$\varphi(X_i) = \sum_{j=1}^{p} \alpha_j \varphi(X_{i-j}) + \varphi(\epsilon_i). \tag{4.4}$$

Se usa el conjunto de coeficientes $\boldsymbol{\lambda}$ para el modelo autorregresivo en $\mathcal{X}$, y un conjunto de coeficientes $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_p]^\top$ para el modelo autorregresivo en $\mathcal{H}$.

Para estimar los parámetros $\boldsymbol{\alpha}$ en el espacio transformado, los autores siguen un procedimiento similar a las ecuaciones de Yule-Walker, pero en lugar de calcular las covarianzas entre las variables aleatorias X_i y X_{i-k}, calculan productos internos entre $\varphi(X_i)$, y $\varphi(X_{i-k})$ en el espacio de Hilbert. Con los supuestos de independencia, el sistema de ecuaciones de Yule-Walker queda dado por

$$\langle \varphi(X_i), \varphi(X_{i-k})\rangle = \sum_{j=1}^{p} \alpha_j \langle \varphi(X_{i-j}), \varphi(X_{i-k})\rangle, \quad \text{para } k = 1, \ldots, p. \tag{4.5}$$

En la literatura, estos productos internos son reemplazados por funciones kernel (el truco del kernel) [39]. Ahora, si se tiene un conjunto de observaciones para el proceso aleatorio de tiempo discreto $\{x_i\}_{i=1}^{N_X}$, el siguiente conjunto de ecuaciones puede ser usado para calcular $\boldsymbol{\alpha}$,

$$k(x_i, x_{i-k}) = \sum_{j=1}^{p} \alpha_j k(x_{i-j}, x_{i-k}), \quad \text{para } k = 1, \ldots, p. \tag{4.6}$$

Dado que los valores $k(x_i, x_{i-j})$, y $k(x_{i-j}, x_{i-k})$ dependen de las observaciones en una serie de tiempo particular y asumiendo que el proceso aleatorio discreto es estacionario, los autores en [23] proponen el siguiente conjunto de ecuaciones para obtener una estimación de $\boldsymbol{\alpha}$

$$\mathbb{E}[k(x_i, x_{i-k})] = \sum_{j=1}^{p} \alpha_j \mathbb{E}[k(x_{i-j}, x_{i-k})], \quad \text{para } k = 1, \ldots, p, \tag{4.7}$$

donde $\mathbb{E}[k(x_i, x_{i-k})]$ y $\mathbb{E}[k(x_{i-j}, x_{i-k})]$ son estimadas sobre un conjunto de muestras disponibles. Este método lo llamaremos modelo autorregresivo kernel (MAK).

Una de nuestras contribuciones es el embebimiento del modelo autorregresivo en un TP-RKHS por medio del mapeo de distribuciones conjuntas $\mathbb{P}(X_i, X_{i-k})$, y $\mathbb{P}(X_{i-j}, X_{i-k})$ a puntos en $\mathcal{H}_1 \otimes \mathcal{H}_2$. Los embebimientos se realizan mediante los operadores de covarianza cruzada, en lugar de productos internos.

Teorema 4.1.3. *Considere el modelo definido en la Ecuación 4.4, y los operadores de covarianza cruzada definidos como*

$$\mathcal{C}_{X_{i-j}X_{i-k}} = \mathbb{E}_{X_{i-j},X_{i-k}}[\varphi(X_{i-j}) \otimes \phi(X_{i-k})] \quad y \quad \mathcal{C}_{X_iX_{i-k}} = \mathbb{E}_{X_i,X_{i-k}}[\varphi(X_i) \otimes \phi(X_{i-k})],$$

entonces

$$\mathcal{C}_{X_iX_{i-k}} = \sum_{j=1}^{p} \alpha_j \mathcal{C}_{X_{i-j}X_{i-k}}, \qquad \textit{para } k = 1, \cdots, p, \tag{4.8}$$

son las Ecuaciones de Yule-Walker en un TP-RKHS.

Prueba Si aplicamos un producto tensor con $\phi(X_{i-k})$, en ambos lados de la Ecuación (4.4), y la esperanza, obtenemos

$$\begin{aligned}\mathbb{E}_{X_i,X_{i-k}}[\varphi(X_i) \otimes \phi(X_{i-k})] = &\sum_{j=1}^{p} \alpha_j \mathbb{E}_{X_{i-j},X_{i-k}}[\varphi(X_{i-j}) \otimes \phi(X_{i-k})] \\ &+ \mathbb{E}_{\epsilon_i X_{i-k}}[\varphi(\epsilon_i) \otimes \phi(X_{i-k})], \quad \text{para } k = 1, \cdots, p.\end{aligned} \tag{4.9}$$

Asumiendo que $\phi(X_{i-k})$, y $\varphi(\epsilon_i)$ son variables que no están correlacionadas, entonces la expresión anterior se reduce a

$$\mathcal{C}_{X_iX_{i-k}} = \sum_{j=1}^{p} \alpha_j \mathcal{C}_{X_{i-j}X_{i-k}}, \quad \text{para } k = 1, \cdots, p. \ \square \tag{4.10}$$

En este trabajo, nos referimos a este método como el método embebimiento kernel (MEK).

4.2 Estimación de los parámetros de un proceso autorregresivo en un TP-RKHS

En esta sección, proporcionamos dos métodos para estimar los parámetros $\boldsymbol{\alpha}$ en el modelo autorregresivo de la Ecuación (4.8). Para este propósito, usamos el estimador de covarianza cruzada y su norma, como en la Ecuación (2.6). Estos métodos los resumimos en los siguientes teoremas:

Teorema 4.2.1. *Sea $\mathcal{C}_{X_iX_{i-k}}$ definido como en (4.8) y considere $\{(x_i^l, x_{i-j}^l)\}_{l=1}^{N_{xy}}$, para $j = 1, 2, \cdots, p$, diferentes conjuntos de muestras tomadas i.i.d de las distribuciones $\mathbb{P}(X_i, X_{i-j})$, entonces el estimador de $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_p]$ está dado por*

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{A}\boldsymbol{\alpha} - \mathbf{b}\|_2^2. \tag{4.11}$$

Además, si $\left(\mathbf{A}^\top\mathbf{A}\right)$ *es invertible, entonces*

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{A}\boldsymbol{\alpha} - \mathbf{b}\|_2^2 = \left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{b}, \tag{4.12}$$

donde $\mathbf{A} \in \mathbb{R}^{p\times p}$ *con entradas* $\{\mathrm{tr}(\widehat{\mathbf{K}}_i^\top\widehat{\mathbf{K}}_j)\}_{i=1,j=1}^{p,p}$, $\mathbf{b} \in \mathbb{R}^{p\times 1}$ *con entradas* $\{\mathrm{tr}(\mathbf{H}^\top\widehat{\mathbf{K}}_i)\}_{i=1}^{p}$, $\mathbf{H} \in \mathbb{R}^{N_{xy}p\times N_{xy}}$ *es una matriz de bloques con bloques dados por* $\{\mathbf{H}_k\}_{k=1}^{p}$ *y* $\widehat{\mathbf{K}}_i \in \mathbb{R}^{N_{xy}p\times N_{xy}}$ *es una matriz de bloques con bloques dados por* $\{\mathbf{K}_{k,i}\}_{k=1}^{p}$, *con* $\mathbf{H}_k = \mathbf{H}_{i-k,i} = \boldsymbol{\Upsilon}_{i-k}^\top\boldsymbol{\Phi}_i$ *and* $\mathbf{K}_{k,i} = \boldsymbol{\Upsilon}_k^\top\boldsymbol{\Upsilon}_i$.

Este teorema es probado en el Apéndice B.

Teorema 4.2.2. *Sea* $\mathcal{C}_{X_iX_{i-k}}$ *definido como en* (4.8) *y considere* $\{(x_i^l, x_{i-j}^l)\}_{l=1}^{N_{xy}}$, *para* $j = 1, 2, \cdots, p$, *diferentes conjuntos de muestras tomadas i.i.d de las distribuciones* $\mathbb{P}(X_i, X_{i-j})$, *entonces el estimador* $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_p]$ *es dado por* $(\mathbf{A}_t^\top\mathbf{A}_t)\widehat{\boldsymbol{\alpha}} = \mathbf{A}_t^\top\mathbf{b}_t$ *donde* $b_k = \mathrm{tr}\left((\mathbf{L}\mathbf{K}_{k,k}\right)$, *y* $a_{j,k} = \mathrm{tr}\left(\mathbf{K}_{j,k}\mathbf{H}_k\right)$ *para* $k = 1, 2, \cdots, p$, *con* $\mathbf{L} = \boldsymbol{\Phi}_i^\top\boldsymbol{\Phi}_i$, $\mathbf{K}_{k,i} = \boldsymbol{\Upsilon}_k^\top\boldsymbol{\Upsilon}_i$ *y* $\mathbf{H}_k = \boldsymbol{\Phi}_i^\top\boldsymbol{\Upsilon}_{i-k}$. *Además si* $(\mathbf{A}_t^\top\mathbf{A}_t)$ *es invertible, entonces*

$$\widehat{\boldsymbol{\alpha}} = \mathbf{A}_t^\top\mathbf{b}_t(\mathbf{A}_t^\top\mathbf{A}_t)^{-1}.$$

Prueba Si hacemos el producto interior de los términos de la Ecuación B.7

$$\boldsymbol{\Upsilon}_{i-k}^\top\boldsymbol{\Phi}_i = \sum_{j=1}^{p}\alpha_j\boldsymbol{\Upsilon}_{i-k}^\top\boldsymbol{\Upsilon}_{i-j}, \tag{4.13}$$

con $\boldsymbol{\Phi}_i\boldsymbol{\Upsilon}_{i-k}^\top$, obtenemos

$$\mathrm{tr}\left(\left(\boldsymbol{\Phi}_i\boldsymbol{\Upsilon}_{i-k}^\top\right)^\top\left(\boldsymbol{\Phi}_i\boldsymbol{\Upsilon}_{i-k}^\top\right)\right) = \sum_{j=1}^{p}\alpha_j\,\mathrm{tr}\left(\left(\boldsymbol{\Upsilon}_{i-j}\boldsymbol{\Upsilon}_{i-k}^\top\right)^\top\left(\boldsymbol{\Phi}_i\boldsymbol{\Upsilon}_{i-k}^\top\right)\right). \tag{4.14}$$

Aplicando la propiedad de la transpuesta de una matriz y la propiedad de la traza de un producto de matrices, obtenemos

$$\mathrm{tr}\left(\left(\boldsymbol{\Phi}_i^\top\boldsymbol{\Phi}\right)\left(\boldsymbol{\Upsilon}_{i-k}^\top\boldsymbol{\Upsilon}_{i-k}\right)\right) = \sum_{j=1}^{p}\alpha_j\,\mathrm{tr}\left(\left(\boldsymbol{\Upsilon}_{i-j}^\top\boldsymbol{\Upsilon}_{i-k}\right)\left(\boldsymbol{\Phi}_i^\top\boldsymbol{\Upsilon}_{i-k}\right)\right). \tag{4.15}$$

Si definimos $\mathbf{L} = \boldsymbol{\Phi}_i^\top\boldsymbol{\Phi}_i$, $\mathbf{H}_k = \boldsymbol{\Phi}_i^\top\boldsymbol{\Upsilon}_{i-k}$ y $\mathbf{K}_{j,k} = \boldsymbol{\Upsilon}_{i-j}^\top\boldsymbol{\Upsilon}_{i-k}$ para $k = 1, 2, \cdots, p$, entonces la Ecuación (4.15) puede escribirse como

$$\mathrm{tr}\left(\mathbf{L}\mathbf{K}_{k,k}\right) = \sum_{j=1}^{p}\alpha_j\,\mathrm{tr}\left(\mathbf{K}_{j,k}\mathbf{H}_k\right) \text{ para } k = 1, 2, \cdots, p. \tag{4.16}$$

Luego la solución de la Ecuación (4.16) es dada por $\widehat{\boldsymbol{\alpha}} = (\mathbf{A}_t^\top\mathbf{A}_t)^{-1}\mathbf{A}_t^\top\mathbf{b}_t$ donde $b_k = \mathrm{tr}\left((\mathbf{L}\mathbf{K}_{k,k}\right)$ y $a_{j,k} = \mathrm{tr}\left(\mathbf{K}_{j,k}\mathbf{H}_k\right)$. □

Note que si definimos la transformación lineal $\mathrm{T} : \mathcal{M}_{N_{xy}\times N_{xy}} \longrightarrow \mathbb{R}$ como $\mathrm{T}(\mathbf{K}) = \mathbb{E}(\mathrm{diag}(\mathbf{K}))$ donde $\mathcal{M}_{N_{xy}\times N_{xy}}$ es el conjunto de todas las matrices de orden $N_{xy} \times N_{xy}$ y si aplicamos T a la Ecuación (B.8),

$$\mathbf{H}_{i-k,i} = \sum_{j=1}^{p} \alpha_j \mathbf{K}_{i-k,i-j}, \tag{4.17}$$

entonces obtenemos

$$\mathbb{E}(\mathrm{diag}(\mathbf{H}_k)) = \sum_{r=1}^{N_{xy}} h(x_{i-k}^r, x_i^r)\mathbb{P}(X_{i-k} = x_{i-k}^r,\ X_i = x_i^r) = \mathbb{E}(h(X_{i-k},\ X_i)),$$

y

$$\mathbb{E}(\mathrm{diag}(\mathbf{K}_{jk})) = \sum_{r=1}^{N_{xy}} k(x_{i-j}^r, x_{i-k}^r)\mathbb{P}(X_{i-j} = x_{i-j}^r,\ X_{i-k} = x_{i-k}^r) = \mathbb{E}(k(X_{i-j},\ X_{i-k})),$$

en consecuencia

$$\mathbb{E}(h(X_{i-k},\ X_i)) = \sum_{j=1}^{p} \alpha_j \mathbb{E}(k(X_{i-j},\ X_{i-k})) \quad \text{para} \quad k = 1, 2, \cdots, p. \tag{4.18}$$

La Ecuación (4.18) es obtenida en [23]. Esto significa que el método presentado en [23] es un caso especial de nuestro método.

4.3 Predicción de series de tiempo usando el problema de la pre-imagen

En esta sección, usamos el método de la pre-imagen para pronosticar un nuevo valor x_i^* usando los valores estimados $\boldsymbol{\alpha}$ y los p valores anteriores de una serie de tiempo. Por ahora, nuestro método expresado en el Teorema 4.2.1 nos permite hacer predicciones en el espacio TP-RKHS por medio de la ecuación

$$\tau_i^* = \sum_{j=1}^{p} \alpha_j \varphi(x_{i-j}) \otimes \phi(x), \tag{4.19}$$

donde los valores para $\{\alpha_j\}_{j=1}^p$ han sido estimados como se explicó en la sección 4.2. Nos gustaría volver a mapear el valor de τ_i^* al espacio de entrada, para obtener la predicción x_i^*. En la literatura de los métodos kernel, este problema se conoce como el problema de la pre-imagen [18], y es un problema mal planteado debido a la mayor dimensionalidad del espacio de Hilbert, lo que significa que los puntos transformados por τ_i^* pueden no tener una correspondencia x_i^* tal que $\varphi(x_i^*) = \tau_i^*$. El problema de la pre-imagen consiste en resolver el siguiente problema de optimización.

Teorema 4.3.1. *Si x_i^* minimiza la expresión*

$$x_i^* = \arg\min_x f(x) = \arg\min_x \left\| \sum_{j=1}^{p} \alpha_j \varphi(x_{i-j}) \otimes \phi(x) - \varphi(x) \otimes \phi(x) \right\|^2_{\mathcal{H}_1 \otimes \mathcal{H}_2},$$

entonces

$$x_{i+1}^* = \frac{\sum_{j=1}^{p} \alpha_j k(x_{i-j}, x_i^*) x_{i-j}}{\sum_{k=1}^{p} \alpha_k k(x_{i-k}, x_i^*)}. \tag{4.20}$$

Este teorema es probado en el Apéndice B.

4.4 Experimentos del modelo autorregresivo basado en un TP-RKHS

En esta sección, proporcionamos información para la evaluación experimental que se realiza en este trabajo. Describimos los conjuntos de datos que utilizamos y el procedimiento que seguimos para validar los resultados.

4.4.1 Descripción de las bases de datos

En este trabajo, utilizamos cuatro series de tiempo con el propósito de evaluar el desempeño de nuestro método. Las primeras dos bases de datos pertenecen a la librería de series de tiempo (TSDL), ver [20]. Las dos últimas bases de datos fueron generadas por el autor de este trabajo.

Rotación de la tierra. Con este nombre, nos referimos a los cambios anuales en la rotación de tierra, disponible en [20]. La serie de tiempo contiene 150 muestras. Utilizamos las primeras 130 muestras para los experimentos.

CO_2. Esta base de datos corresponde a las medidas mensuales de CO_2 en partes por millón desde el Observatorio de Mauna Loa. La base de datos contiene 192 muestras. Para los experimentos se utilizan las primeras 150 muestras.

MG_{30}. Esta serie de tiempo, se refiere a las serie de tiempo obtenida de la solución de la ecuación diferencial no lineal Mackey-Glass dada como

$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t-\tau)}{1+x(t-\tau)},$$

con $\tau = 30$ [23]. Esta serie de tiempo exhibe dinámica caótica. Generamos una serie de tiempo de longitud 600. Para los experimentos, se utilizan las primeras 400 muestras.

El atractor de Lorenz. Esta base de dato se refiere a la solución del sistema de tres ecuaciones diferenciales ordinarias acopladas donde r, a y b son constantes

$$\begin{aligned}\frac{dx(t)}{dt} &= -ax + ay \\ \frac{dy(t)}{dt} &= -xz + rx - y \\ \frac{dz(t)}{dt} &= xy - bz.\end{aligned}$$

Establecemos los valores para los parámetros $a = 10$, $r = 28$ y $b = 8/3$. Para estos valores, las tres series de tiempo multivariadas $(x(t), y(t), z(t))$ muestran dinámica caótica. Generamos 500 muestras por dimensión de salida, y usamos las primeras 400 muestras para los experimentos. Realizamos predicción sobre las tres series de tiempo $x(t)$, $y(t)$ y $z(t)$, tratándolas como independientes unas de otras.

4.4.2 Validación del modelo AR basado en un TP-RKHS

La validación del método que proponemos en este trabajo se realiza mediante la predicción un paso hacia adelante sobre cada una de las series de tiempo descritas anteriormente. Para realizar la predicción un paso hacia adelante, utilizamos marcos de deslizamientos de $w+1$ muestras, donde las primeras w muestras son usadas para estimar los parámetros de los modelos utilizados, incluyendo el orden p del modelo autorregresivo y la última muestra adicional se utiliza para la validación. Los marcos de deslizamientos son organizados consecutivamente, con una superposición de w muestras. Los datos de entrenamiento se utilizan para ajustar los parámetros de cada uno de los modelos utilizados para la comparación, incluyendo el orden del modelo autorregresivo. Para el orden del modelo, evaluamos los valores de p de uno a cinco. Calculamos el error cuadrático medio sobre las muestras de validación.

Los modelos utilizados en los experimentos y que son comparados con el método que proponemos son:

Modelo autorregresivo lineal (LAR). Este modelo se describe en la Definición 4.1.1. Los coeficientes $\boldsymbol{\lambda}$ para el modelo lineal AR son estimados usando las ecuaciones de Yule-Waker. Dentro de cada marco de longitud w, usamos una ventana deslizante de tamaño $w/2+1$, donde las primeras $w/2$ muestras son usadas para calcular $\boldsymbol{\lambda}$, y la última muestra se utiliza para realizar la predicción un paso hacia adelante para diferentes valores de p. Las ventanas deslizantes se organizan consecutivamente con una superposición de $w/2$ muestras. Los resultados obtenidos en la predicción un paso hacia adelante dentro del marco de longitud w, se utilizan para seleccionar el valor de p, el cual es seleccionado como el valor que ocurrió con mayor frecuencia ofreciendo el mejor rendimiento en la predicción. Una vez que se ha seleccionado el valor de p, nosotros calculamos otra vez los valores de $\boldsymbol{\lambda}$ usando todos los puntos que están en w y usamos este nuevo $\boldsymbol{\lambda}$ para realizar la predicción un paso hacia adelante en el paso de tiempo $w+1$.

Modelo autorregresivo kernel (MAK). El modelo MAK es propuesto en [23] y es presentado en la Sección 4.1. La predicción de este modelo se hace resolviendo el

problema de la pre-imagen como se explica en [18]. Los valores de ℓ y p son elegidos de la siguiente manera: dentro de cada marco de longitud w, se generan ventanas deslizantes de tamaño $w/2+1$. Las ventanas deslizantes se organizan consecutivamente con una superposición de $w/2$ muestras. Las primeras $w/2$ muestras son usadas para estimar los valores de $\boldsymbol{\alpha}$ mediante la solución del sistema de ecuaciones de la expresión (4.7). Luego se utiliza el dato en el paso del tiempo $w/2+1$ para seleccionar los mejores valores de ℓ y p, que en promedio dentro de la ventana de longitud w producen el menor error. Luego se utiliza una grilla de valores de ℓ tomando una grilla de porcentajes, ℓ_p, de la mediana de los datos de entrenamiento en la ventana de tamaño $w/2$. Los porcentajes que se consideran son 0.01, 0.01, 0.5, 1, 2 o 5 de la mediana de los datos del entrenamiento en la ventana de longitud $w/2$. Una vez se selecciona el valor de ℓ_p, se calcula un nuevo valor para ℓ como el porcentaje ℓ_p de la mediana de los datos del entrenamiento en la ventana de tamaño w. Después de elegir ℓ y p, se utilizan todos los datos de entranamiento dentro de la ventana w para buscar un nuevo conjunto de coeficientes de $\boldsymbol{\alpha}$ y por último, se realiza un pronóstico en el paso del tiempo $w+1$ mediante la solución del problema de la pre-imagen.

Método embebimiento de distribuciones en un RKHS o también llamado método de embebimiento de kernel (KEM). Se implementa el método que se describe en el Capítulo 2. También se utiliza un kernel Gaussiano. Los valores para ℓ y p utilizados para hacer predicción un paso hacia adelante en el tiempo $w+1$ se calculan como sigue: dentro de cada marco de de longitud w, utilizamos una ventana deslizante de la longitud $w/2+1$. Las ventanas deslizantes se organizan consecutivamente con una superposición de $w/2$ muestras. Las primeras $w/2$ muestras son usadas para estimar los coeficientes $\boldsymbol{\lambda}$ mediante la solución de la ecuación (B.11). Para seleccionar ℓ y p, se sigue el mismo procedimiento al utilizado para MAK.

Procesos Gaussianos (GP). Seguimos el modelo propuesto en [22], en el cual la variable aleatoria del proceso en el tiempo X_n puede ser descrita usando la siguiente ecuación

$$X_n = f(X_{n-1}, \ldots, X_{n-p}) + \epsilon, \tag{4.21}$$

donde $\epsilon \sim \mathcal{N}(0, \sigma^2)$, y $f(\mathbf{x})$ se supone que sigue apriori un proceso Gaussiano $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x})))$, con función de covarianza $k(\mathbf{x}, \mathbf{x})$. Para la función de covarianza, se utiliza el kernel Gaussiano como en la Ecuación (B.19). El parámetro de la función de covarianza ℓ y el parámetro σ para el modelo de probabilidad se estiman maximizando el logaritmo de la probabilidad marginal, usando un procedimiento de gradiente conjugado. Utilizamos GPmat Toolbox [1] para todas las rutinas relacionadas con los procesos Gaussianos. Para seleccionar el valor de p, usamos un marco deslizante de longitud $w/2+1$, dentro del marco de longitud w. Las ventanas deslizantes de longitud $w/2+1$ se realizan como en los métodos anteriores. Se utilizan varios puntos de datos $w/2$ para estimar los hiperparámetros del proceso Gaussiano, y el punto de dato en el paso de tiempo $w/2+1$ se utiliza para la validación cruzada del valor de p. El valor para p se elige como el que en promedio (dentro del marco de tamaño w) conduce a los errores más bajos. Una vez que el valor para p se ha elegido, utilizamos nuevamente las muestras w para entrenar un nuevo GP. Esta nueva GP ajustada se usa para pronosticar

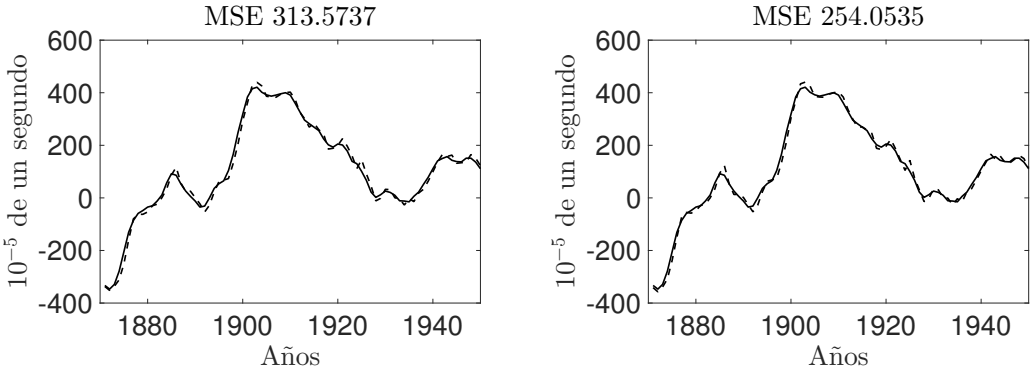
[1] Disponible en https://github.com/SheffieldML/GPmat

el punto de datos en el paso de tiempo $w+1$.

Redes Neuronales (NN). En este modelo se utiliza una red neuronal con una capa oculta y para el aprendizaje se utiliza un mapeo similar al de la Ecuación (4.21). Para elegir el número de neuronas n_h de la capa oculta, y el valor de p, usamos un procedimiento similar al de los métodos anteriores: dentro del marco de longitud w, generamos ventanas deslizantes de longitud $w/2+1$, de forma similar a como se deslizaron en los otros enfoques. Los primeros puntos de datos $w/2$ se utilizan para ajustar los pesos de la red neuronal, y el punto de datos en el paso de tiempo $w/2+1$ se utiliza para elegir el valor de n_h, y el valor de p. Estos valores se eligen como los que en promedio, dentro del marco de longitud w, conduce al error más bajo. Permitimos que la cantidad de neuronas en la capa oculta sea cualquiera de los siguientes valores: 5, 10, 15, 20, 25 o 30. Para las rutinas de las redes neuronales, utilizamos la caja de herramientas redes neuronales para MATLAB, con todas las configuraciones predeterminadas, excepto para el número de neuronas en la capa oculta [52].

4.5 Análisis de los resultados

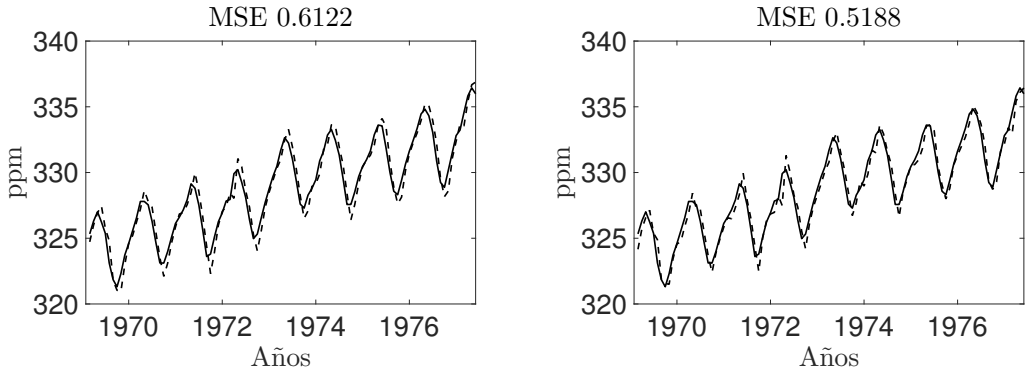
En esta sección se compara el rendimiento de los diferentes métodos para la predicción a corto plazo sobre cada una de las cuatro series de tiempo descritas en la Sección 4.4. Las Figuras 4.1, 4.2, 4.3, y 4.4 muestran el rendimiento del modelo autorregresivo kernel y el modelo autorregresivo basado en el método embebimiento de distribuciones de probabilidad en un RKHS sobre las cuatro series de tiempo descritas en Sección 4.4. El error cuadrático medio (MSE) para la predicción de un paso hacia delante, es mostrado como el título en cada figura.



(a) Rotación de la tierra usando MAK (b) Rotación de la tierra usando MEK

Figura 4.1: Predicción un paso hacia adelante sobre el conjunto de datos rotación de la tierra dado por el método propuesto por Kallas et. al. en [23] y el método basado en embebimiento de distribuciones de probabilidad en un RKHS propuesto en este trabajo. Las líneas sólidas son los datos de prueba, las líneas punteadas son las predicciones dadas por los métodos. El título de cada figura muestra el MSE entre los datos de prueba y las predicciones de salida.

La Figura 4.1 muestra los resultados de la predicción un paso hacia adelante para la serie de tiempo rotación de la tierra. Para este ejemplo, se usa ventanas deslizantes de longitud 51. Los primeras 50 observaciones de cada ventana deslizante fueron usadas para el entrenamiento, y la predicción fue realizada para el paso del tiempo 51 de cada ventana deslizante. Note que ambos métodos kernel (figures 4.1(a) y 4.1(b)) son capaces de seguir la serie de tiempo original incluso los primeros pasos de tiempo. Los valores para el MSE muestran que el método basado en embebimiento de distribuciones de probabilidad en un RKHS ofrece el mejor rendimiento en comparación con el método autorregresivo de kernel propuesto por Kallas.

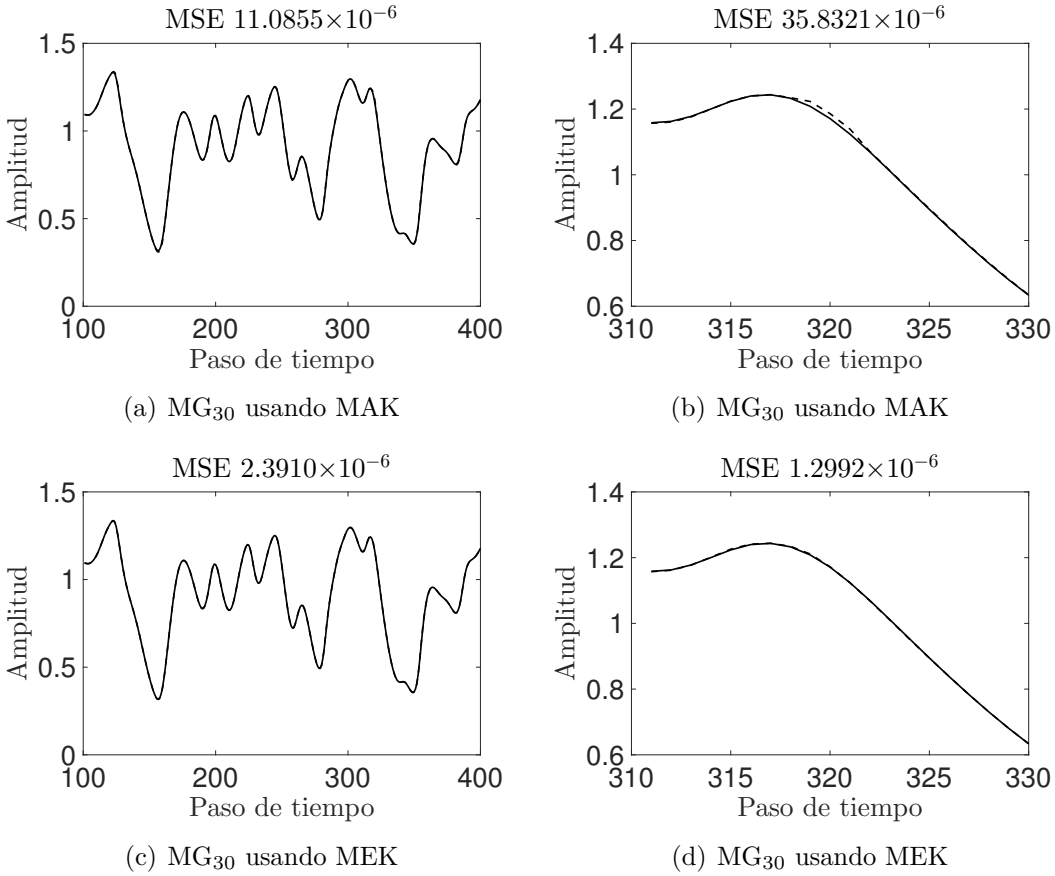


(a) CO_2 usando MAK (b) CO_2 usando MEK

Figura 4.2: Predicción un paso hacia adelante sobre el conjunto de datos CO_2, dado por el método propuesto por Kallas et. al. en [23] y el método basado en embebimiento de distribuciones de probabilidad en un RKHS propuesto en este trabajo. Las líneas sólidas son los datos de prueba, las líneas punteadas son las predicciones dadas por los métodos. El título de cada figura muestra el MSE entre los datos de prueba y las predicciones de salida.

La Figura 4.2 muestra los resultados de predicciones de un paso hacia adelante sobre el conjunto de datos CO_2 para los modelos MAK y MEK. Como en el ejemplo anterior, se utilizan ventanas deslizantes de longitud de 51 muestras, donde las primeras 50 muestras en cada ventana se utilizan para encontrar los parámetros de los modelos, y la última muestra (número 51) se utiliza para probar la capacidad de predicción de los métodos de prueba. De la serie de tiempo CO_2 que esta originalmente disponible, se usan las primeras 150 muestras para evaluar el rendimiento de las predicciones en varios puntos de la serie tiempo. Luego, se usa un marco de ventana de 51 puntos, el MSE para la predicción se calcula para 100 muestras de las series de tiempo. Se puede apreciar cómo el modelo MEK es capaz de seguir más de cerca los bajos y altos valores de pico de la serie de tiempo, cuando se compara con el modelo MAK. Esto puede explicarse por el hecho de que el método embebimiento de distribuciones de probabilidad en un RKHS es capaz de tener en cuenta la forma como varía la serie de tiempo, que para el MAK se pierde cuando se promedian los datos. Los valores MSE (que aparecen en el título de cada figura) indican que el modelo KEM supera al modelo MAK. A partir de los experimentos con la base de datos CO_2 se pudo observar que con respecto a los valores de p, el modelo lineal AR funciona mejor con un valor de $p = 2$, o $p = 5$, en su mayoría. El KAM consistentemente funcionaba mejor con $p = 2$, y el KEM con $p = 5$. Con respecto a los valores elegidos para ℓ en los métodos kernel, en solo 4 de los 80 ensayos, ambos métodos eligieron diferentes valores de ancho de banda. Los valores

MSE indican que el MEK supera al método lineal AR y al método MAK.



(a) MG_{30} usando MAK

(b) MG_{30} usando MAK

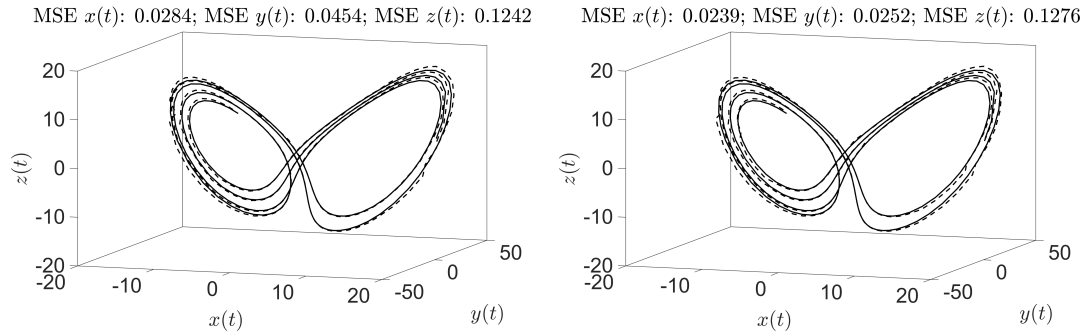
(c) MG_{30} usando MEK

(d) MG_{30} usando MEK

Figura 4.3: La predicción un paso hacia adelante sobre la base de datos MG_{30} dado por, el método propuesto por Kallas in [23] y el método embebimiento de distribuciones de probabilidad en un RKHS propuesto en este trabajo. Las líneas sólidas son los datos de prueba, las líneas punteadas son las predicciones dadas por los métodos. Las figuras 4.3(a) y 4.3(c) muestran los resultados para la serie de tiempo MG_{30}. Las figuras 4.3(b) y 4.3(d) muestran resultados para la serie de tiempo MG_{30} dentro de un período más corto de tiempo, entre pasos de tiempo 311 y 330. El título de cada figura muestra el MSE entre los datos de prueba y las predicciones de salida.

La Figura 4.3 muestra la predicción un paso hacia adelante para la serie de tiempo caótica de Mackey-Glass. Para esta serie de tiempo, utilizamos ventanas deslizantes de longitud 101. Las primeras 100 muestras de la ventana deslizante son usadas para entrenar el modelo, y la muestra 101 se utiliza para la predicción un paso hacia adelante. El rendimiento en predicción de los modelos MAK y MEK son similares. Las Figuras 4.3(b) y 4.3(d) muestran los resultados para la serie de tiempo MG_{30} dentro de un período de tiempo más corto, entre los pasos de tiempo 311 y 330. El experimento muestra que el MSE obtenido por el modelo MEK es menor que el obtenido por los modelos MAK y LAR. Con respecto a los valores de p, el modelo LAR favorecido es para el valor de $p = 5$ (250 sobre los 300 ensayos). Los modelos MAK y MEK que predominan, utilizan altos valores de p: 70 por $p = 4$ y 163 para $p = 5$ para el MAK; y 43 para $p = 4$ y 257 para $p = 5$ para el MEK. En contraste con los experimentos anteriores, esta vez los métodos kernel sólo seleccionan el mismo valor ℓ en 76 casos

de 300. En términos de la media del MSE en los 300 ensayos, el experimento muestra que ambos métodos kernel superan al método LAR. El MSE obtenido por el MEK es menor que el obtenido por MAK.



(a) Lorenz usando MAK (b) Lorenz usando MEK

Figura 4.4: Predicción un paso hacia adelante sobre la base de datos de Lorenz dado por el método propuesto por [23] y el método embebimiento de distribuciones de probabilidad en un RKHS propuesto en este trabajo. Las líneas sólidas son los datos de prueba, las líneas punteadas son las predicciones dadas por los métodos. El título de cada figura muestra la media del MSE entre los datos de prueba y la predicción de salida.

La Figura 4.4 muestra los resultados de predicción sobre la base de datos de Lorenz. Como se explicó antes, la predicción se realiza sobre cada componente $(x(t), y(t), z(t))$ de la serie de tiempo 3D, de manera independiente. Para cada una de las tres series de tiempo, utilizamos ventanas deslizantes de longitud 101, donde la predicción se realiza sobre el último paso de tiempo de cada marco. Se evalúa el desempeño de la predicción sobre 300 sucesivos marcos, todos ellos de longitud 101. Con respecto al orden p para los diferentes modelos, el modelo LAR seleccionado es para $p = 2$ de casi 22% de las veces y $p = 5$ de casi 75% de las veces. El modelo MAK elegido es para $p = 2$ de casi el 70% de todos los ensayos y $p = 3$ de casi 22% todas las repeticiones. Por último, el modelo MEK elegido es para $p = 2$ de casi 42% de veces, $p = 4$ de casi 23% de las veces y $p = 5$ aproximadamente 30% de los ensayos. En cuanto a los experimentos anteriores, los métodos kernel superan al modelo lineal AR. También se observa de los experimentos, que el error en la predicción del modelo MAK para $z(t)$ es menor que el error en la predicción para MEK. Sin embargo, en promedio el modelo MEK supera al modelo MAK.

La Tabla 4.1 muestra un resumen del MSE obtenido por el modelo lineal AR, el modelo autorregresivo kernel y el modelo basado en embebimiento de distribuciones de probabilidad en un RKHS para los cuatro conjuntos de datos. A partir de la Tabla 4.1 se puede observar que el método de embebimiento de distribuciones de probabilidad en un RKHS tiene mejor rendimiento, salvo la componente $z(t)$ de la serie de tiempo de Lorenz.

La Tabla 4.2 muestra el rendimiento de las redes neuronales, procesos Gaussianos y el método kernel autorregresivo comparados con el rendimiento del método de embebimiento de distribuciones de probabilidad en un RKHS propuesto en este trabajo. La Tabla 4.2 muestra que los métodos basados en kernels, MAK, MEK Y GP, tienen mejor rendimiento en la predicción que el método NN. El proceso Gaussiano supera

Tabla 4.1: Error cuadrático medio para los datos de prueba y las predicciones de salidas, dadas por el modelo AR, el modelo MAK y el modelo MEK. Los valores de los MSE para el MG_{30} deben ser multiplicados por 10^{-6}.

Database	Lineal AR	MAK	MEK
Earthrot	689.3491	313.5737	**254.0535**
CO_2	0.6572	0.6122	**0.5188**
MG_{30}	372.0770	11.0855	**2.3910**
Lorenz $x(t)$	0.3051	0.0284	**0.0239**
Lorenz $y(t)$	0.9905	0.0454	**0.0252**
Lorenz $z(t)$	0.4371	**0.1242**	0.1276

Tabla 4.2: El MSE para los datos de prueba y las predicciones de salidas, dadas por una red neuronal (NN), un regresor proceso Gaussiano (GP), el modelo kernel autorregresivo propuesto por Kallas et. al. en [23] (MAK) y el modelo aotorregresivo basado en embebimiento de distribuciones de probabilidad en un RKHS propuesto en este documento (MEK). Los valores de los MSE para el MG_{30} deben ser multiplicados por 10^{-6}.

Database	NN	GP	MAK	MEK
Earthrot	827.8469	570.1689	182.3038	**134.2564**
CO_2	0.6027	0.4631	0.4107	**0.3177**
MG_{30}	41.4519	2.0991	6.3064	**1.1052**
Lorenz $x(t)$	0.0595	0.0118	0.0088	**0.0037**
Lorenz $y(t)$	0.1114	0.0129	0.0146	**0.0038**
Lorenz $z(t)$	0.2041	0.0263	0.0211	**0.0140**

al método MAK para la serie de tiempo MG_{30} y la componente de $y(t)$ de la serie de tiempo de Lorenz. El método MEK muestra mejor rendimiento sobre todos los otros modelos que compiten [52].

4.6 Resumen y comentarios del Capítulo 4

En este capítulo hemos aplicado el método embebimiento de distribuciones de probabilidad conjunta en un TP-RKHS de un proceso autorregresivo de orden p a través del operador de covarianza cruzada. La estimación del modelo se hizo mediante un sistema de ecuaciones de Yule-Walker definido en el espacio TP-RKHS. Las predicciones se realizaron resolviendo un problema de pre-imagen, para lo cual se desarrolló un algoritmo de punto fijo. Los resultados experimentales muestran que el método propuesto aquí supera a varias versiones no lineales del modelo autoregresivo, en la tarea de hacer predicción un paso hacia adelante de series de tiempo.

Capítulo 5

Conclusiones y trabajos futuros

5.1 Conclusiones

En este trabajo, hemos presentado nuevas métricas entre distribuciones de probabilidad y entre HMMs, usando el método embebimiento de distribuciones de probabilidad en un RKHS. Estas métricas se obtienen, bajo la suposición de que los kernels característicos son el kernel Gaussianos y el kernel Laplaciano. Además, se supone que las distribuciones de probabilidad son estimadas usando el estimador de Parzen, y en el caso de los HMMs se supone que estos son estacionarios y sus distribuciones de probabilidad son mezclas de Gaussianas.

Del mismo modo, hemos introducido el método embebimiento de distribuciones de probabilidad conjunta en un TP-RKHS por medio de un proceso autorregresivo de orden p usando los operadores de la covarianza cruzada. La estimación de los parámetros del modelo se hace a través de un sistema de ecuaciones de Yule-Walker, y la estimación empírica de los operadores de covarianza cruzada. Las predicciones en el espacio de entrada se realizan mediante la solución de un problema de pre-imagen, para lo cual se desarrolla un algoritmo de punto fijo. Los resultados experimentales muestran cómo este método aplicado a modelos autorregresivos, produce excelentes predicciones un paso hacia adelante en series de tiempo. Estas predicciones son mejores que las obtenidas por el método de MAK, el método lineal, el método de redes neuronales y el método de procesos Gaussianos.

5.2 Trabajos futuros

El rendimiento en clasificación de patrones usando métricas basadas en el método embebimiento de distibuciones de probabilidad en un RKHS depende fundamentalmente del kernel característico empleado en la métrica. En este trabajo, hemos usado kernels característicos suaves como el kernel Gaussiano y el kernel Laplaciano. La pregunta que nos hacemos después de realizar este trabajo de investigación es: ¿cómo será el rendimiento en clasificación de patrones usando métricas basadas en el método embebimiento de distribuciones de probabilidad en un RKHS usando otros kernels característicos?. Un trabajo futuro, sería proponer nuevas métricas entre distribuciones

de probabilidad usando este método para otros kernels caracteríticos distintos al kernel Laplaciano y al kernel Gaussiano.

En este trabajo de investigación, desarrollamos métricas entre HMMs usando el método embebimiento de distribuciones de probabilidad en un RKHS para hacer clasificación de series de tiempo. En la construcción de estas métricas tuvimos en cuenta que los HMMs son estacionarios y que los kernels característicos son suaves, como por ejemplo el kernel Gaussiano y el kernel Laplaciano. Las métricas que desarrollamos en este trabajo tuvieron buen rendimento en clasificacion de series de tiempo, especialmente cuando las series de tiempo son suaves y estacionarias. En series de tiempo no tan suaves el rendimiento de las métricas propuestas no fue tan bueno. Una posible solución al mejoramiento del rendimiento en clasificación de este tipo de series de tiempo, sería construir una métrica basada en los RKHS para HMMs con otros tipos de kernels no tan suaves y construir una métrica cuando los HMMs no son estacionarios, es decir, cuando la matriz de transición se toma en cuenta para la definición de la métrica.

Los resultados que obtuvimos en la predicción de series de tiempo estacionarias, usando el modelo autorregresivo y el método embebimiento de distribuciones en un RKHS fueron más que interesantes. Sin embargo, estos resultados se podrian mejorar si se supone que las series de tiempo siguen un modelo distinto al modelo autorregresivo como por ejemplo los modelos ARIMA, ARMA y martingalas. Una extensión de esta línea de trabajo sería aplicar la métodologia de los embebimientos en un RKHS a los modelos ARIMA, ARMA y martingalas. Nosotros consideramos la estimación y predicción de un modelo autorregresivo no lineal, donde los valores de los coeficientes $\{\alpha_j\}_{j=1}^{p}$ en otro futuro trabajo tendrá que considerarse como operadores lineales más generales.

Referencias

[1] C. R. Baker. Joint measures and cross-covariance operators. *American Mathematical Society*, 9:273–289, 1973.

[2] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distribution. *Calcutta Math. Soc*, 35(04):99–109, 1943.

[3] M. Bicego, V. Murino, and Mário Figueiredo. A sequential pruning strategy for the selection of the number of states in hidden Markov models. *Pattern Recognition Letters*, 24(9):1395–1407, 2003.

[4] C. M. Bishop. *Pattern Recognition And Machine Learning (Information Science And Statistics)*. Springer, 2007.

[5] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer series in statistics. Springer, 1st edition, 2005.

[6] S. Chakraborty. Bayesian binary kernel probit model for microarray based cancer classification and gene selection. *Computational Statistics and Data Analysis*, 53 (12):4198 – 4209, 2009.

[7] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The UCR time series classification archive, 2015.

[8] Y. Chen, J. Ye, and J. Li. A distance for HMMs based on aggregated Wasserstein metric and state registration. *ArXiv e-prints*, 2016.

[9] K. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in neural information processing systems*, pages 3608–3616, 2014.

[10] P. Corazza. Introduction to metric-preserving functions. *The American mathematical monthly*, 106(4):309–323, 1999.

[11] T. Cover and J. Thomas. Elements of information theory. *New Yor: Wiley*, 2nd-Ed, 1991.

[12] K. Fukumizu, F. Bach, and M. Jordan. Kernel dimensionality reduction for supervised learning. *Advances in Neural Information Processing Systems*, 16:81, 2004.

[13] K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14(1): 3753–3783, 2013.

[14] A. Gretton, O. Bousquet, A. Smola, and B. Schöllkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings algotithmic learning theory*, pages 63–77. Springer-Verlag, 2005.

[15] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

[16] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[17] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[18] P. Honeine and C. Richard. Preimage problem in kernel-based machine learning. *IEEE Signal Processing Magazine*, 28:73–88, 2011.

[19] M. Houda et al. Probability metrics and the stability of stochastic programs with recourse. *Bulletin of the Czech Econometric Society*, 9(17):65–77, 2002.

[20] R. Hyndman. Time series data library. http://data.is/TSDLdemo.

[21] Y. Inamura. Estimating continuous time transition matrices from discretely observed data. *Bank of Japan working paper series*, E07(06):1–40, 2006.

[22] B. Banko J. Kocijan, A. Girard and R. Murray-Smith. Dynamic systems identification with, Gaussian processes. *Mathematical and Computer modelling of Dynamical Systems*, 11(4):411–424, 2005.

[23] M. Kallas, P. Honeine, C. Francis, and H. Amoud. Kernel autoregressive models using Yule Walker equations. *Signal Processing*, 93:3053–3061, 2013.

[24] R. Kumar and C. V. Jawahar. Kernel approach to autoregressive modeling. In *13th National Conference on Comunications (NCC) Kanpur, India*, 2007.

[25] R. Lee. Modeling demographic relationhips: Analysis of forescast funtions for autralian births. *American Statistical Association*, 76:782–792, 1981.

[26] V. V. Leigh. The statistics of variation. *Variation: A central concept in biology*, pages 29–48, 2005.

[27] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[28] S. Mirasgedis, Y. Sarafidis, E. G. Poulou, F. Karagiannis D. P. Lalas, M. Moschovits, and D. Papakonstantinou. Models for mid-term electricity demand forecasting incorporating weather influences. *Elsevier: Energy*, 31:208–227, 2006.

[29] K. Muandet, K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Kernel mean estimation and Stein effect. In *Proceedings of the 31st International Conference on Machine Learning, W & CP 32 (1)*, pages 10–18. JMLR, 2014.

[30] A. Muller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

[31] K. P. Murphy. *Machine Learning: A Probabilistic Perspective (Adaptive Computation And Machine Learning Series)*. The MIT Press, 2012.

[32] R. M. Neal and E. G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

[33] O. Nelles. *Nonlinear System Identification: from Classical approaches to Neural Networks and fuzzy models*. Springer, first edition edition, 2001.

[34] M. Piccardi and Ó. Pérez. Hidden Markov models with kernel density estimation of emission probabilities and their use in activity recognition. IEEE Computer Society, 2007. ISBN 1-4244-1179-3.

[35] L. Ralaivola and F. d'Alché Buc. Time series filtering, smoothing and learning using the kernel kalman filter. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 3, pages 1449–1454. IEEE, 2005.

[36] A. Ramdas and L. Wehbe. Stein shrinkage for cross-covariance operators and kernel independence testing. *ArXiv e-prints*, 2014.

[37] C. Richard, J. C. M. Bermudez, and P. Honeine. Online prediction of time series data with kernels. *Signal Processing, IEEE Transactions on*, 57(3):1058–1067, 2009.

[38] Tobias Rydén. Estimating the order of hidden Markov models. *Statistics: A Journal of Theoretical and Applied Statistics*, 26(4):345–354, 1995.

[39] B. Schölkopf and A. J. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[40] L. Scrucca et al. Nonparametric kernel smoothing methods. The SM library in XLISP-STAT. *Journal of Statistical Software*, 6(7):1–49, 2001.

[41] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[42] K. S. Shanmugan and A. M. Breipohi. *Random Signals: Detection, Estimation and Data Analysis*. Wiley, first edition, 1988.

[43] Hideaki Shimazaki and Shigeru Shinomoto. Kernel bandwidth optimization in spike rate estimation. *J. Comput. Neurosci.*, 29:171–182, 2010.

[44] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31, 2007.

[45] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert embeddings of conditional distributions with applications to dynamical systems. In *26th Annual International Conference on Machine Learning Montreal-Canada*, pages 961–968, 2009.

[46] L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola. Hilbert space embeddings of hidden Markov. In *27th Annual International Conference on Machine Learning, Haifa-Israel*, 2010.

[47] L. Song, A. Gretton, and K. Fukumizu. Embedding of conditional distributions. *IEEE. Signal Processing Magazine*, 30:98–111, 2013.

[48] C. Soto and M. González.

[49] B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

[50] E. Trentin and M. Gori. A survey of hybrid ANN and HMM models for automatic speech recognition. *Neurocomputing*, 37(1-4):91–126, 2001.

[51] R. S. Tsay. *Analysis of Financial Time Series*. Wiley, third edition, 2010.

[52] E. A. Valencia and M. A. Álvarez. Short-term time series prediction using Hilbert space embeddings of autoregressive processes. *Neurocomputing*, 266:595 – 605, 2017.

[53] R. W. Vallin et al. Continuity and differentiability aspects of metric preserving functions. *Real Analysis Exchange*, 25(2):849–868, 1999.

[54] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

[55] L. Yang, B. K. Widjaja, and R. Prasad. Application of hidden Markov models for signature verification. *Pattern Recognition*, 28(2):161–170, 1995.

[56] J. Zeng, j. Duan, and C. Wu. A new distance measure for hidden Markov models. *Expert Systems with Applications*, 37(2):1550–1555, 2010.

[57] J. Zeng, L. Xie, U. Kruger, J. Yu, J. Sha, and X. Fu. Process monitoring based on Kullback Leibler divergence. In *Control Conference (ECC), 2013 European*, pages 416–421. IEEE, 2013.

[58] S. Zhong and J. Ghosh. A unified framework for model-based clustering. *Journal of Machine Learning Research*, 4:1001–1037, 2003.

[59] Carlos D Zuluaga, Edgar A Valencia, Mauricio A Álvarez, and Álvaro A Orozco. A Parzen-based distance between probability measures as an alternative of summary statistics in approximate Bayesian computation. In *International Conference on Image Analysis and Processing*, pages 50–61. Springer, 2015.

Appendix A

Pruebas de teoremas sobre métricas entre distribuciones de probabilidad usando el método embebimiento de distribuciones de probabilidad en un RKHS

Este capítulo contiene las pruebas de los teoremas más importantes sobre métricas entre distribuciones de probabilidad usando el método embebimiento de distribuciones de probabilidad en un RKHS. A continuación se demuestran los Teoremas 3.1.1 y 3.1.5 de la Sección 3.1.

Teorema 3.1.1 Si el kernel $k(\mathbf{x},\mathbf{x}')$ es un kernel Gaussiano con parámetro $\mathbf{\Sigma}$, y los estimadores $\widehat{p}(\mathbf{x})$, y $\widehat{q}(\mathbf{y})$ son estimados como

$$\widehat{p}(\mathbf{x}) = \frac{1}{N_x}\sum_{i=1}^{N_x} \frac{1}{\left(2\pi\sigma_p^2\right)^{D/2}} \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma_p^2}\right) = \frac{1}{N_x}\sum_{i=1}^{N_x} \mathcal{N}(\mathbf{x}|\mathbf{x}_i,\mathbf{\Sigma}_p), \quad \mathbf{\Sigma}_p = \mathbf{I}\sigma_p^2, \tag{A.1}$$

$$\widehat{q}(\mathbf{y}) = \frac{1}{N_y}\sum_{j=1}^{N_y} \frac{1}{\left(2\pi\sigma_q^2\right)^{D/2}} \exp\left(-\frac{\|\mathbf{y}-\mathbf{y}_j\|^2}{2\sigma_q^2}\right) = \frac{1}{N_y}\sum_{j=1}^{N_y} \mathcal{N}(\mathbf{y}|\mathbf{y}_j,\mathbf{\Sigma}_q), \quad \mathbf{\Sigma}_q = \mathbf{I}\sigma_q^2, \tag{A.2}$$

respectivamente, donde $\mathbf{I}$ es la matriz identidad, σ_p y σ_q son los anchos de banda de kernel, y D es la dimensión del espacio de entrada, entonces

$$\begin{aligned}\widehat{\gamma_k^2}(\mathbb{P},\mathbb{Q}) = &\frac{1}{N_x^2}\sum_{i,j=1}^{N_x} \hat{k}\left(\mathbf{x}_i,\mathbf{x}_j;2\mathbf{\Sigma}_p\right) + \frac{1}{N_y^2}\sum_{i,j=1}^{N_y} \hat{k}\left(\mathbf{y}_i,\mathbf{y}_j;2\mathbf{\Sigma}_q\right) \\ &- \frac{2}{N_x N_y}\sum_{i,j=1}^{N_x,N_y} \hat{k}\left(\mathbf{x}_i,\mathbf{y}_j;\mathbf{\Sigma}_p+\mathbf{\Sigma}_q\right),\end{aligned} \tag{A.3}$$

donde

$$\hat{k}\left(\mathbf{x},\mathbf{x}';\mathbf{S}\right)=\frac{|\boldsymbol{\Sigma}|^{1/2}}{|\boldsymbol{\Sigma}+\mathbf{S}|^{1/2}}\exp\left(-\frac{\left(\mathbf{x}-\mathbf{x}'\right)^{\top}\left(\boldsymbol{\Sigma}+\mathbf{S}\right)^{-1}\left(\mathbf{x}-\mathbf{x}'\right)}{2}\right).$$

Prueba. Primero calculamos la integral $I_2=\int_{\mathcal{X}}\int_{\mathcal{X}}k(\mathbf{x},\mathbf{y})p(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y}$, con $\mathcal{X}=\mathbb{R}^D$. Supongamos que $\widehat{p}(\mathbf{x})$ y $\widehat{q}(\mathbf{y})$ son estimadores de $p(\mathbf{x})$ y $q(\mathbf{y})$ respectivamente, dados por

$$\widehat{p}(\mathbf{x})=\frac{1}{N_x}\sum_{i=1}^{N_x}\mathcal{N}(\mathbf{x}|\mathbf{x}_i,\boldsymbol{\Sigma}_p),\quad \boldsymbol{\Sigma}_p=\mathbf{I}\sigma_p^2, \tag{A.4}$$

$$\widehat{q}(\mathbf{y})=\frac{1}{N_y}\sum_{j=1}^{N_y}\mathcal{N}(\mathbf{y}|\mathbf{y}_j,\boldsymbol{\Sigma}_q),\quad \boldsymbol{\Sigma}_q=\mathbf{I}\sigma_q^2, \tag{A.5}$$

y sea

$$\frac{1}{|\boldsymbol{\Sigma}|^{1/2}(2\pi)^{D/2}}k(\mathbf{x},\mathbf{y})=\mathcal{N}(\mathbf{x}-\mathbf{x}|\mathbf{0},\boldsymbol{\Sigma}).$$

Usando las propiedades de la distribución Gaussiana

$$\mathcal{N}(\mathbf{x}-\mathbf{y}|\mathbf{0},\boldsymbol{\Sigma})=\mathcal{N}(\mathbf{x}|\mathbf{y},\boldsymbol{\Sigma})=\mathcal{N}(\mathbf{y}|\mathbf{x},\boldsymbol{\Sigma}),$$

obtenemos

$$\begin{aligned}
I_2 &= \int_{\mathcal{X}}\int_{\mathcal{X}}\frac{|\boldsymbol{\Sigma}|^{1/2}}{(2\pi)^{-D/2}}\mathcal{N}(\mathbf{x}-\mathbf{y}|\mathbf{0},\boldsymbol{\Sigma})\left(\frac{1}{N_x}\sum_{i=1}^{N_x}\mathcal{N}(\mathbf{x}|\mathbf{x}_i,\boldsymbol{\Sigma}_p)\right)\left(\frac{1}{N_y}\sum_{j=1}^{N_y}\mathcal{N}(\mathbf{y}|\mathbf{y}_j,\boldsymbol{\Sigma}_q)\right)d\mathbf{x}d\mathbf{y}\\
&= \frac{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}{N_xN_y}\sum_{j=1}^{N_y}\sum_{i=1}^{N_x}\int_{\mathcal{X}}\int_{\mathcal{X}}\mathcal{N}(\mathbf{y}|\mathbf{x},\boldsymbol{\Sigma})\mathcal{N}(\mathbf{x}|\mathbf{x}_i,\boldsymbol{\Sigma}_p)\mathcal{N}(\mathbf{y}|\mathbf{y}_j,\boldsymbol{\Sigma}_q)d\mathbf{x}d\mathbf{y} \qquad (A.6)\\
&= \frac{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}{N_xN_y}\sum_{j=1}^{N_y}\sum_{i=1}^{N_x}\int_{\mathcal{X}}\left(\int_{\mathcal{X}}\mathcal{N}(\mathbf{y}|\mathbf{x},\boldsymbol{\Sigma})\mathcal{N}(\mathbf{x}|\mathbf{x}_i,\boldsymbol{\Sigma}_p)d\mathbf{x}\right)\mathcal{N}(\mathbf{y}|\mathbf{y}_j,\boldsymbol{\Sigma}_q)d\mathbf{y}\\
&= \frac{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}{N_xN_y}\sum_{j=1}^{N_y}\sum_{i=1}^{N_x}\int_{\mathcal{X}}\mathcal{N}(\mathbf{y}|\mathbf{x}_c,\boldsymbol{\Sigma}_c)\mathcal{N}(\mathbf{y}|\mathbf{x}_j,\boldsymbol{\Sigma}_q)d\mathbf{y},
\end{aligned}$$

donde $\mathbf{x}_c=\mathbf{x}_i$ y $\boldsymbol{\Sigma}_c=\boldsymbol{\Sigma}_p+\boldsymbol{\Sigma}$. Ahora, si usamos la identidad del producto entre dos distribuciones Gaussianas

$$\mathcal{N}(\mathbf{y}|\mathbf{x}_c,\boldsymbol{\Sigma}_c)\mathcal{N}(\mathbf{y}|\mathbf{x}_j,\boldsymbol{\Sigma}_q) = \mathcal{N}(\mathbf{y}|\mathbf{y}_T,\boldsymbol{\Sigma}_T)\mathcal{N}(\mathbf{x}_c|\mathbf{x}_j,\boldsymbol{\Sigma}_c+\boldsymbol{\Sigma}_q), \tag{A.7}$$

donde

$$\mathbf{y}_T=\left(\boldsymbol{\Sigma}_c^{-1}+\boldsymbol{\Sigma}_p^{-1}\right)^{-1}\left(\mathbf{x}_c\boldsymbol{\Sigma}_c^{-1}+\mathbf{x}_j\boldsymbol{\Sigma}_p^{-1}\right),$$

y

$$\mathbf{\Sigma}_T = \left(\mathbf{\Sigma}_c^{-1} + \mathbf{\Sigma}_p^{-1}\right)^{-1} = \left(\mathbf{\Sigma}_p^{-1} + \mathbf{\Sigma}^{-1} + \mathbf{\Sigma}_p^{-1}\right)^{-1},$$

entonces se obtiene la expresión

$$I_2 = \frac{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}}{N_x N_y} \sum_{j=1}^{N_y}\sum_{i=1}^{N_x} \mathcal{N}(\mathbf{x}_c|\mathbf{x}_j, \mathbf{\Sigma}_c + \mathbf{\Sigma}_q) \int_{\mathcal{X}} \mathcal{N}(\mathbf{y}|\mathbf{y}_T, \Sigma_T) d\mathbf{y}. \quad (A.8)$$

Ahora, dado que $\int_X \mathcal{N}(\mathbf{y}|\mathbf{y}_T, \mathbf{\Sigma}_T)d\mathbf{y} = 1$, en consecuencia

$$I_2 = \frac{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}{N_x N_y} \sum_{j=1}^{N_y}\sum_{i=1}^{N_x} \mathcal{N}(\mathbf{x}_i|\mathbf{x}_j, \mathbf{\Sigma} + \mathbf{\Sigma}_p + \mathbf{\Sigma}_q). \quad (A.9)$$

El mismo procedimiento usado para obtener I_2 se usa para obtener I_1 y I_3, esto es

$$I_1 = \int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{x},\mathbf{z})p(\mathbf{x})p(\mathbf{z})d\mathbf{x}d\mathbf{z} = \frac{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}}{N_x^2} \sum_{j=1}^{N_x}\sum_{i=1}^{N_x} \mathcal{N}(\mathbf{x}_i|\mathbf{x}_j, \mathbf{\Sigma} + 2\mathbf{\Sigma}_p), \quad (A.10)$$

y

$$I_3 = \int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{y},\mathbf{z})q(\mathbf{y})q(\mathbf{x})d\mathbf{y}d\mathbf{z} = \frac{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}}{N_y^2} \sum_{j=1}^{N_y}\sum_{i=1}^{N_y} \mathcal{N}(\mathbf{y}_i|\mathbf{y}_j, \mathbf{\Sigma} + 2\mathbf{\Sigma}_q). \quad (A.11)$$

Si reemplazamos I_1, I_2 y I_3 en la Ecuación (2.11), obtenemos el estimador $\widehat{\gamma_k^2}(\mathbb{P},\mathbb{Q})_{Parzen}$ de $\gamma_k^2(\mathbb{P},\mathbb{Q})$

$$\begin{aligned}
\widehat{\gamma_k^2}(\mathbb{P},\mathbb{Q})_{Parzen} &= \frac{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}}{N_x^2} \sum_{j=1}^{N_x}\sum_{i=1}^{N_x} \mathcal{N}(\mathbf{x}_i|\mathbf{x}_j, \mathbf{\Sigma} + 2\mathbf{\Sigma}_p) \\
&+ \frac{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}}{N_y^2} \sum_{j=1}^{N_y}\sum_{i=1}^{N_y} \mathcal{N}(\mathbf{y}_i|\mathbf{y}_j, \mathbf{\Sigma} + 2\mathbf{\Sigma}_q) \quad (A.12) \\
&-2 \frac{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}}{N_x N_y} \sum_{j=1}^{N_x}\sum_{i=1}^{N_y} \mathcal{N}(\mathbf{x}_i|\mathbf{y}_j, \mathbf{\Sigma} + \mathbf{\Sigma}_p + \mathbf{\Sigma}_q) \\
&= \frac{1}{N_x^2}\sum_{i,j=1}^{N_x} \hat{k}\left(\mathbf{x}_i, \mathbf{x}_j; 2\mathbf{\Sigma}_p\right) + \frac{1}{N_y^2}\sum_{i,j=1}^{N_y} \hat{k}\left(\mathbf{y}_i, \mathbf{y}_j; 2\mathbf{\Sigma}_q\right) \quad (A.13) \\
&- \frac{2}{N_x N_y}\sum_{i,j=1}^{N_x,N_y} \hat{k}\left(\mathbf{x}_i, \mathbf{y}_j; \mathbf{\Sigma}_p + \mathbf{\Sigma}_q\right),
\end{aligned}$$

donde

$$\hat{k}\left(\mathbf{x}, \mathbf{x}'; \mathbf{S}\right) = \frac{|\mathbf{\Sigma}|^{1/2}}{|\mathbf{\Sigma} + \mathbf{S}|^{1/2}} \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')^\top (\mathbf{\Sigma} + \mathbf{S})^{-1} (\mathbf{x} - \mathbf{x}')}{2}\right). \square$$

Teorema 3.1.5. Si el kernel $k(x,x';\ell)$ es un kernel Laplaciano con parámetro ℓ y

$$\widehat{p}(x) = \frac{1}{\sigma_p\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right),$$
$$\widehat{q}(y) = \frac{1}{\sigma_q\sqrt{2\pi}}\exp\left(-\frac{(y-\mu_q)^2}{2\sigma_q^2}\right),$$

son estimadores de $p(x)$ y $q(y)$ respectivamente, donde los parámetros ℓ, μ_p, μ_q, σ_p, $\sigma_q \in \mathbb{R}$ y los valores de entrada $x, y \in \mathbb{R}$, entonces un nuevo estimador de la métrica entre las distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$ es obtenido a partir de la expresión (2.11)

$$\begin{aligned}\widehat{\gamma_k^2}(\mathbb{P},\mathbb{Q}) &= \frac{1}{2\pi\sigma_p^2}\left(\mathcal{I}_1(\mu_p,\sigma_p,\ell)+\mathcal{I}_2(\mu_p,\sigma_p,\ell)\right)\\ &+ \frac{1}{2\pi\sigma_q^2}\left(\mathcal{I}_1(\mu_q,\sigma_q,\ell)+\mathcal{I}_2(\mu_q,\sigma_q,\ell)\right)\\ &\quad \frac{-1}{\pi\sigma_p\sigma_q}\left(\mathcal{I}_1(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell)+\mathcal{I}_2(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell)\right), \qquad (A.14)\end{aligned}$$

donde

$$\mathcal{I}_1(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell) = f_1(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell)\frac{\sigma_p\sqrt{\pi}}{\sigma_q}\left(1-\operatorname{erf}\left(\frac{d_1(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell)\,\sigma_p}{2\sigma_q\sqrt{\sigma_q^2+\sigma_p^2}}\right)\right), \quad (A.15)$$

$$d_1 = d_1(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell) = \frac{\sigma_q}{\sqrt{2}\ell^2\sigma_p^2}\left(2\ell^2(\mu_q-\mu_p)+\sigma_p^2+\sigma_q^2\right), \qquad (A.16)$$

$$\begin{aligned}f_1(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell) &= \sigma_q^2\sqrt{\pi}\exp\left(\frac{\left(2\ell^2\mu_q+\sigma_q^2\right)^2}{8\ell^4\sigma_q^2}\left(1-\frac{\sigma_q^2}{\sigma_p^2}\right)\right) \qquad (A.17)\\ &\times \exp\left(-\frac{\left(2\ell^2\mu_q+\sigma_q^2\right)}{2\ell^2}\left(\frac{1}{2\ell^2}-\frac{\mu_p}{\sigma_p}\right)-\frac{1}{2}\left(\frac{\mu_q^2}{\sigma_q^2}+\frac{\mu_p^2}{\sigma_p^2}\right)+\frac{\sigma_p^2 d_1^2}{4\sigma_q^2}\right),\end{aligned}$$

$$\mathcal{I}_2(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell) = f_2(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell)\frac{\sigma_q\sqrt{\pi}}{\sigma_p}\left(1-\operatorname{erf}\left(\frac{d_2(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell)\,\sigma_q^2}{2\sigma_p\sqrt{\sigma_p^2+\sigma_q^2}}\right)\right),$$

$$d_2 = d_2(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell) = \frac{\sigma_p}{\sqrt{2}\ell^2\sigma_q^2}\left(2\ell^2(\mu_p-\mu_q)+\sigma_q^2+\sigma_p^2\right), \qquad (A.18)$$

$$f_2\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right) = \sigma_p^2\sqrt{\pi}\exp\left(\frac{\left(2\ell^2\mu_p+\sigma_p^2\right)^2}{8\ell^4\sigma_p^2}\left(1-\frac{\sigma_p^2}{\sigma_q^2}\right)\right) \tag{A.19}$$
$$\times \exp\left(-\frac{\left(2\ell^2\mu_p+\sigma_p^2\right)}{2\ell^2}\left(\frac{1}{2\ell^2}-\frac{\mu_q}{\sigma_q^2}\right)-\frac{1}{2}\left(\frac{\mu_p^2}{\sigma_p^2}+\frac{\mu_q^2}{\sigma_q^2}\right)+\frac{\sigma_q^2 d_2^2}{4\sigma_p^2}\right).$$

Prueba. Sea

$$\exp\left(\frac{-|x-y|}{2\ell^2}\right)=\begin{cases}\exp\left(\frac{-(x-y)}{2\ell^2}\right) & si \quad x\geq y\\ \exp\left(\frac{(x-y)}{2\ell^2}\right) & si \quad x<y,\end{cases} \tag{A.20}$$

por lo tanto

$$\begin{aligned}\mathcal{I}\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right) &= \frac{1}{2\pi\sigma_p\sigma_q}\int_{-\infty}^{\infty}\int_{-\infty}^{x}\exp\left(\frac{-(x-y)}{2\ell^2}\right)\exp\left(\frac{-(x-\mu_p)^2}{2\sigma_p^2}\right)\\ &\times \exp\left(\frac{-(y-\mu_q)^2}{2\sigma_q^2}\right)dydx+\frac{1}{2\pi\sigma_p\sigma_q}\int_{-\infty}^{\infty}\int_{-\infty}^{y}\exp\left(\frac{(x-y)}{2\ell^2}\right)\\ &\times \exp\left(\frac{-\left(x-\mu_p^2\right)^2}{2\sigma_p^2}\right)\exp\left(\frac{-(y-\mu_q)^2}{2\sigma_q^2}\right)dxdy \qquad (A.21)\\ &= \frac{1}{2\pi\sigma_p\sigma_q}\left(\mathcal{I}_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right)+\mathcal{I}_2\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right)\right).\end{aligned}$$

Calculemos $\mathcal{I}_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right)$.

$$\begin{aligned}\mathcal{I}_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{x}\exp\left(\frac{-(x-y)}{2\ell^2}\right)\exp\left(\frac{-(x-\mu_p)^2}{2\sigma_p^2}\right)\exp\left(\frac{-(y-\mu_q)^2}{2\sigma_q^2}\right)dydx\\ &= \int_{-\infty}^{\infty}\exp\left(\frac{-(x-\mu_p)^2}{2\sigma_p^2}\right)\left[\int_{-\infty}^{x}\exp\left(\frac{-(x-y)}{2\ell^2}\right)\exp\left(\frac{-(y-\mu_q)^2}{2\sigma_q^2}\right)dy\right]dx,\end{aligned}$$

donde

$$\begin{aligned}\int_{-\infty}^{x}\exp\left(\frac{-(x-y)}{2\ell^2}\right)\exp\left(\frac{-(y-\mu_q)^2}{2\sigma_q^2}\right)dy &= \int_{-\infty}^{x}\exp\left(\frac{-y^2+2y\mu_q-\mu_q^2}{2\sigma_q^2}-\frac{(x-y)}{2\ell^2}\right)dy\\ &= \int_{-\infty}^{x}\exp\left(-\frac{y^2}{2\sigma_q^2}+y\left(\frac{\mu_q}{\sigma_q^2}+\frac{1}{2\ell^2}\right)-\frac{\mu_q^2}{2\sigma_q^2}-\frac{x}{2\ell^2}\right)dy\\ &= \exp\left(-\frac{\mu_q^2}{2\sigma_q^2}-\frac{x}{2\ell^2}\right)\\ &\times \int_{-\infty}^{x}\exp\left(-\frac{y^2}{2\sigma_q^2}+y\left(\frac{\mu_q}{\sigma_q^2}+\frac{1}{2l^2}\right)\right)dy. \qquad (A.22)\end{aligned}$$

Sea $a = \frac{1}{2\sigma_q^2}$ y $b = \frac{\mu_q}{\sigma_q^2} + \frac{1}{2\ell^2} = \frac{2\ell^2\mu^q + \sigma_q^2}{2l^2\sigma_q^2}$, luego

$$\begin{aligned}\int_{-\infty}^{x} \exp\left(\frac{-(x-y)}{2\ell^2}\right)\exp\left(\frac{-(y-\mu_q)^2}{2\sigma_q^2}\right)dy &= \exp\left(-\frac{\mu_q^2}{2\sigma_q^2} - \frac{x}{2\ell^2}\right)\int_{-\infty}^{x}\exp\left(-ay^2+yb\right)dy \\ &= \exp\left(-\frac{\mu_q^2}{2\sigma_q^2} - \frac{x}{2\ell^2} + \frac{b^2}{4a}\right) \\ &\times \int_{-\infty}^{x}\exp\left(-a\left(y-\frac{b}{2a}\right)^2\right)dy.\end{aligned}$$

Ahora, supongamos que $u = \sqrt{a}\left(y - \frac{b}{2a}\right)$, esto implica que $y = \frac{u}{\sqrt{a}} + \frac{b}{2a}$ y $dy = \frac{du}{\sqrt{a}}$. Si $y = x$, entonces $u = a\left(x - \frac{b}{2a}\right)$, y si $y = -\infty$ entonces $u = -\infty$. Análogamente, si $z = \sqrt{a}\left(x - \frac{b}{2a}\right)$, esto implica que $x = \frac{z}{\sqrt{a}} + \frac{b}{2a}$ y $dx = \frac{dz}{\sqrt{a}}$, por lo tanto

$$\begin{aligned}\int_{-\infty}^{x} \exp\left(\frac{-(x-y)}{2\ell^2}\right)\exp\left(\frac{-(y-\mu_q)^2}{2\sigma_q^2}\right)dy &= \exp\left(-\frac{\mu_q^2}{2\sigma_q^2} - \frac{\left(\frac{z}{\sqrt{a}}+\frac{b}{2a}\right)}{2\ell^2} + \frac{b^2}{4a}\right) \\ &\times \int_{-\infty}^{z}\exp\left(-u^2\right)\frac{du}{\sqrt{a}} \\ &= \exp\left(-\frac{\mu_q^2}{2\sigma_q^2} - \frac{\left(\frac{z}{\sqrt{a}}+\frac{b}{2a}\right)}{2l^2} + \frac{b^2}{4a}\right) \\ &\times \frac{\sqrt{\pi}}{2\sqrt{a}}\left(1+\operatorname{erf}(z)\right), \qquad (A.23)\end{aligned}$$

donde

$$\int_{-\infty}^{z}\exp\left(-u^2\right)du = \frac{\sqrt{\pi}}{2}\left(1+\operatorname{erf}(z)\right) \quad \text{and} \quad \operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z \exp\left(-u^2\right)du,$$

por consiguiente

$$\begin{aligned}\mathcal{I}_1(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell) &= \frac{\sqrt{\pi}}{2a}\int_{-\infty}^{\infty}\exp\left(-\frac{\mu_q^2}{2\sigma_q^2} - \frac{\left(\frac{z}{\sqrt{a}}+\frac{b}{2a}\right)}{2\ell^2} + \frac{b^2}{4a}\right) \\ &\times \exp\left(\frac{-\left(\frac{z}{\sqrt{a}}+\frac{b}{2a}-\mu_p\right)^2}{2\sigma_p^2}\right)\left(1+\operatorname{erf}(z)\right)dz \\ &= \frac{\sqrt{\pi}}{2a}\exp\left(-\frac{b}{4\ell^2 a} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{b^2}{4a} - \frac{(b-2a\mu_p)^2}{8a^2\sigma_p^2}\right) \\ &\times \int_{-\infty}^{\infty}\exp\left(-\frac{z}{2\sqrt{a}\ell^2} - \frac{z^2}{2a\sigma_p^2} - \frac{z\,(b-2a\mu_p)}{2a\sqrt{a}\sigma_p^2}\right) \\ &\times \left(1+\operatorname{erf}(z)\right)dz. \qquad (A.24)\end{aligned}$$

Por otro lado, se tiene

$$\begin{aligned}\exp\left(-\frac{z}{2\sqrt{a}\ell^2}-\frac{z^2}{2a\sigma_p^2}-\frac{z\,(b-2a\mu_p)}{2a\sqrt{a}\sigma_p^2}\right) &= \exp\left(-\frac{z^2}{2a\sigma_p^2}-z\left(\frac{b-2a\mu_p}{2a\sqrt{a}\sigma_p^2}+\frac{1}{2\sqrt{a}\ell^2}\right)\right)\\ &= \exp\left(-cz^2-zd_1\right)=\exp\left(-c\left(z+\frac{d_1}{2c}\right)^2+\frac{d_1^2}{4c}\right),\end{aligned}$$

donde $c=\dfrac{1}{2a\sigma_p^2}=\dfrac{\sigma_q^2}{\sigma_p^2}$ y $d_1=\dfrac{b-2a\mu_p}{2a\sqrt{a}\sigma_p^2}+\dfrac{1}{2\sqrt{a}\ell^2}$, en consecuencia

$$\begin{aligned}\mathcal{I}_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right) &= \frac{\sqrt{\pi}}{2a}\int_{-\infty}^{\infty}\exp\left(-\frac{b}{4\ell^2a}-\frac{\mu_q^2}{2\sigma_q^2}+\frac{b^2}{4a}-\frac{(b-2a\mu_p)^2}{8a^2\sigma_p^2}+\frac{d_1^2}{4c}\right)\\ &\times \exp\left(-c\left(z+\frac{d_1}{2c}\right)^2\right)(1+\operatorname{erf}(z))\,dz\\ &= f_1\int_{-\infty}^{\infty}\exp\left(-c\left(z+\frac{d}{2c}\right)^2\right)(1+\operatorname{erf}(z))\,dz, \qquad (A.25)\end{aligned}$$

donde

$$f_1=f_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right)=\frac{\sqrt{\pi}}{2a}\exp\left(-\frac{b}{4\ell^2a}-\frac{\mu_q^2}{2\sigma_q^2}+\frac{b^2}{4a}-\frac{(b-2a\mu_p)^2}{8a^2\sigma_p^2}+\frac{d_1^2}{4c}\right). \qquad (A.26)$$

Por otra parte, si usamos la identidad

$$\int_{-\infty}^{\infty}\operatorname{erf}(z)\exp\left(-(az+b)^2\right)dx=\frac{-\sqrt{\pi}}{a}\operatorname{erf}\left(\frac{b}{\sqrt{a^2+1}}\right), \qquad (A.27)$$

entonces

$$\int_{-\infty}^{\infty}\operatorname{erf}(z)\exp\left(-\left(\sqrt{c}z+\frac{d_1}{2\sqrt{c}}\right)^2\right)dz=\frac{-\sqrt{\pi}}{\sqrt{c}}\operatorname{erf}\left(\frac{d_1}{2\sqrt{c^2+c}}\right), \qquad (A.28)$$

y

$$\int_{-\infty}^{\infty}\exp\left(-\left(\sqrt{c}z+\frac{d_1}{2\sqrt{c}}\right)^2\right)dz=\frac{\sqrt{\pi}}{\sqrt{c}}. \qquad (A.29)$$

Por consiguiente

$$\mathcal{I}_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right) = f_1\left(\frac{\sqrt{\pi}}{\sqrt{c}}-\frac{\sqrt{\pi}}{\sqrt{c}}\operatorname{erf}\left(\frac{d_1}{2\sqrt{c^2+c}}\right)\right). \qquad (A.30)$$

Ahora, si escribimos d_1 y f_1 en función de los parámetros μ_p, μ_q, σ_p, σ_q y ℓ, entonces

$$d_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right) = \frac{b-2a\mu_p}{2a\sqrt{a}\sigma_p^2}+\frac{1}{2\sqrt{a}\ell^2}=\frac{\sigma_q}{\sqrt{2}\ell^2\sigma_p^2}\left(2\ell^2(\mu_q-\mu_p)+\sigma_p^2+\sigma_q^2\right),$$

$$
\begin{aligned}
f_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right) &= \frac{\sqrt{\pi}}{2a}\exp\left(-\frac{b}{4\ell^2 a}-\frac{\mu_q^2}{2\sigma_q^2}+\frac{b^2}{4a}-\frac{\left(b-2a\mu_p\right)^2}{8a^2\sigma_p^2}+\frac{d_1^2}{4c}\right)\\
&= \sigma_q^2\sqrt{\pi}\exp\left(\frac{\left(2\ell^2\mu_q+\sigma_q^2\right)^2}{8\ell^4\sigma_q^2}\left(1-\frac{\sigma_q^2}{\sigma_p^2}\right)\right) \qquad (A.31)\\
&\times \exp\left(-\frac{\left(2\ell^2\mu_q+\sigma_q^2\right)}{2\ell^2}\left(\frac{1}{2\ell^2}-\frac{\mu_p}{\sigma_p^2}\right)-\frac{1}{2}\left(\frac{\mu_q^2}{\sigma_q^2}+\frac{\mu_p^2}{\sigma_p^2}\right)+\frac{\sigma_p^2 d_1^2}{4\sigma_q^2}\right).
\end{aligned}
$$

Luego

$$
\mathcal{I}_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right)=f_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right)\left(\frac{\sigma_p\sqrt{\pi}}{\sigma_q}-\frac{\sigma_p\sqrt{\pi}}{\sigma_q}\operatorname{erf}\left(\frac{d_1\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right)\sigma_p^2}{2\sigma_q\sqrt{\sigma_q^2+\sigma_p^2}}\right)\right).
$$

Calculemos $\mathcal{I}_2$.

$$
\begin{aligned}
\mathcal{I}_2\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{y}\exp\left(\frac{(x-y)}{2\ell^2}\right)\exp\left(\frac{-\left(x-\mu_p\right)^2}{2\sigma_p^2}\right)\exp\left(\frac{-\left(y-\mu_q\right)^2}{2\sigma_q^2}\right)dxdy\\
&= \int_{-\infty}^{\infty}\exp\left(\frac{-\left(y-\mu_q\right)^2}{2\sigma_q^2}\right)\left[\int_{-\infty}^{y}\exp\left(\frac{(x-y)}{2\ell^2}\right)\exp\left(\frac{-\left(x-\mu_p\right)^2}{2\sigma_p^2}\right)dx\right]dy.
\end{aligned}
$$

Resolvamos la integral

$$
\begin{aligned}
\int_{-\infty}^{y}\exp\left(\frac{(x-y)}{2\ell^2}\right)\exp\left(\frac{-\left(x-\mu_p\right)^2}{2\sigma_p^2}\right)dx &= \int_{-\infty}^{y}\exp\left(\frac{-x^2+2x\mu_p-\mu_p^2}{2\sigma_p^2}+\frac{(x-y)}{2\ell^2}\right)dy\\
&= \int_{-\infty}^{y}\exp\left(-\frac{x^2}{2\sigma_p^2}+x\left(\frac{\mu_p}{\sigma_p^2}+\frac{1}{2\ell^2}\right)-\frac{\mu_p^2}{2\sigma_p^2}-\frac{y}{2\ell^2}\right)dx\\
&= \exp\left(-\frac{\mu_p^2}{2\sigma_p^2}-\frac{y}{2\ell^2}\right) \qquad (A.32)\\
&\times \int_{-\infty}^{y}\exp\left(-\frac{x^2}{2\sigma_p^2}+x\left(\frac{\mu_p}{\sigma_p^2}+\frac{1}{2\ell^2}\right)\right)dy.
\end{aligned}
$$

Sea $a_2=\dfrac{1}{2\sigma_p^2}$ y $b_2=\dfrac{\mu_p}{\sigma_p^2}+\dfrac{1}{2\ell^2}=\dfrac{2\ell^2\mu_p+\sigma_p^2}{2\ell^2\sigma_p^2}$, por lo tanto

$$
\begin{aligned}
\int_{-\infty}^{y}\exp\left(\frac{(x-y)}{2\ell^2}\right)\exp\left(\frac{-\left(x-\mu_p\right)^2}{2\sigma_p^2}\right)dx &= \exp\left(-\frac{\mu_p^2}{2\sigma_p^2}-\frac{y}{2\ell^2}\right)\int_{-\infty}^{y}\exp\left(-a_2x^2+xb_2\right)dx\\
&= \exp\left(-\frac{\mu_p^2}{2\sigma_p^2}-\frac{y}{2\ell^2}+\frac{b_2^2}{4a_2}\right)\\
&\times \int_{-\infty}^{y}\exp\left(-a_2\left(x-\frac{b_2}{2a_2}\right)^2\right)dx.
\end{aligned}
$$

Si $u=\sqrt{a_2}\left(x-\dfrac{b_2}{2a_2}\right)$, entonces $x=\dfrac{u}{\sqrt{a_2}}+\dfrac{b_2}{2a_2}$ y $dx=\dfrac{du}{\sqrt{a_2}}$. Similarmente, si hacemos $x=y$ y $x=-\infty$, esto implica que $u=\sqrt{a_2}\left(y-\dfrac{b_2}{2a_2}\right)$ y $u=-\infty$. Ahora, si $z=\sqrt{a_2}\left(y-\dfrac{b_2}{2a_2}\right)$, entonces $y=\dfrac{z}{\sqrt{a_2}}+\dfrac{b_2}{2a_2}$ y $dy=\dfrac{dz}{\sqrt{a_2}}$. Por consiguiente

$$\begin{aligned}\int_{-\infty}^{y}\exp\left(\frac{(x-y)}{2\ell^2}\right)\exp\left(\frac{-(x-\mu_p)^2}{2\sigma_p^2}\right)dx &= \exp\left(-\frac{\mu_p^2}{2\sigma_p^2}-\frac{\left(\frac{z}{\sqrt{a_2}}+\frac{b_2}{2a_2}\right)}{2\ell^2}+\frac{b_2^2}{4a_2}\right)\\ &\times \int_{-\infty}^{z}\exp\left(-u^2\right)\frac{du}{\sqrt{a_2}}\\ &= \exp\left(-\frac{z}{2\sqrt{a_2}\ell^2}-\frac{\mu_p^2}{2\sigma_p^2}-\frac{b_2}{4\ell^2a_2}+\frac{b_2^2}{4a_2}\right)\\ &\times \frac{\sqrt{\pi}}{2a_2}\left(1+\operatorname{erf}(z)\right)dz, \qquad (A.33)\end{aligned}$$

y

$$\begin{aligned}\mathcal{I}_2\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right) &= \frac{\sqrt{\pi}}{2a_2}\int_{-\infty}^{\infty}\exp\left(-\frac{b_2}{4\ell^2a_2}-\frac{\mu_p^2}{2\sigma_p^2}+\frac{b_2^2}{4a_2}-\frac{z}{2\sqrt{a_2}\ell^2}\right) \qquad (A.34)\\ &\times \exp\left(-\frac{\left(\frac{z}{\sqrt{a_2}}+\frac{b_2}{2a_2}-\mu_q\right)^2}{2\sigma_q^2}\right)\left(1+\operatorname{erf}(z)\right)dz\\ &= \frac{\sqrt{\pi}}{2a_2}\int_{-\infty}^{\infty}\exp\left(-\frac{b_2}{4\ell^2a_2}-\frac{\mu_p^2}{2\sigma_p^2}+\frac{b_2^2}{4a_2}-\frac{\left(b_2-2a_2\mu_q\right)^2}{8a_2^2\sigma_q^2}\right)\\ &\times \exp\left(-\frac{z^2}{2a_2\sigma_q^2}-z\left(\frac{b_2-2a_2\mu_q}{2\sqrt{a_2}a_2\sigma_q^2}+\frac{1}{2\sqrt{a_2}\ell^2}\right)\right)\left(1+\operatorname{erf}(z)\right)dz.\end{aligned}$$

Es decir

$$\begin{aligned}\mathcal{I}_2\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right) &= \frac{\sqrt{\pi}}{2a_2}\int_{-\infty}^{\infty}\exp\left(-\frac{b_2}{4\ell^2a_2}-\frac{\mu_p^2}{2\sigma_p^2}+\frac{b_2^2}{4a_2}-\frac{\left(b_2-2a_2\mu_q\right)^2}{8a_2^2\sigma_q^2}+\frac{d_2^2}{4c_2}\right)\\ &\times \exp\left(-c_2\left(z+\frac{d_2}{2c_2}\right)^2\right)\left(1+\operatorname{erf}(z)\right)dz\\ &= f_2\int_{-\infty}^{\infty}\exp\left(-c_2\left(z+\frac{d_2}{2c_2}\right)^2\right)\left(1+\operatorname{erf}(z)\right)dz, \qquad (A.35)\end{aligned}$$

donde $c_2=\dfrac{1}{2a_2\sigma_q^2}=\dfrac{\sigma_p^2}{\sigma_q^2}$, $d_2=\dfrac{b_2-2a_2\mu_q}{2\sqrt{a_2}a_2\sigma_q^2}+\dfrac{1}{2\sqrt{a_2}l^2}$ y

$$f_2 = f_2\left(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell\right)=\frac{\sqrt{\pi}}{2a_2}\exp\left(-\frac{b_2}{4\ell^2a_2}-\frac{\mu_p^2}{2\sigma_p^2}+\frac{b_2^2}{4a_2}-\frac{\left(b_2-2a_2\mu_q\right)^2}{8a_2^2\sigma_q^2}+\frac{d_2^2}{4c_2}\right).$$

Ahora, si escribimos d_2 y f_2 en función de los parámetros $\mu^{\mathbb{P}}, \mu^{\mathbb{Q}}, \sigma^{\mathbb{P}}, \sigma^{\mathbb{Q}}$ and l, obtenemos las ecuaciones

$$d_2\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right) = \frac{b_2 - 2a_2\mu_q}{2a_2\sqrt{a_2}\sigma_q^2} + \frac{1}{2\sqrt{a_2}\ell^2} = \frac{\sigma_p}{\sqrt{2}\ell^2\sigma_q^2}\left(2\ell^2(\mu_p - \mu_q) + \sigma_q^2 + \sigma_p^2\right),$$

$$\begin{aligned} f_2\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right) &= \frac{\sqrt{\pi}}{2a_2}\exp\left(-\frac{b_2}{4l^2a_2} - \frac{\mu_p^2}{2\sigma_p^2} + \frac{b_2^2}{4a_2} - \frac{\left(b_2 - 2a_2\mu_q\right)^2}{8a_2^2\sigma_q^2} + \frac{d_2^2}{4c_2}\right) \\ &= \sigma_p^2\sqrt{\pi}\exp\left(\frac{\left(2\ell^2\mu_p + \sigma_p^2\right)^2}{8\ell^4\sigma_p^2}\left(1 - \frac{\sigma_p^2}{\sigma_q^2}\right)\right) \qquad \text{(A.36)} \\ &\times \exp\left(-\frac{\left(2\ell^2\mu_p + \sigma_p^2\right)}{2\ell^2}\left(\frac{1}{2\ell^2} - \frac{\mu_q}{\sigma_q^2}\right) - \frac{1}{2}\left(\frac{\mu_p^2}{\sigma_p^2} + \frac{\mu_q^2}{\sigma_q^2}\right) + \frac{\sigma_q^2 d^2}{4\sigma_p^2}\right). \end{aligned}$$

En consecuencia

$$\mathcal{I}_2\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right) = f_2\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right)\left(\frac{\sigma_q\sqrt{\pi}}{\sigma_p} - \frac{\sigma_q\sqrt{\pi}}{\sigma_p}\operatorname{erf}\left(\frac{d_2\left(\mu_p, \mu_q, \sigma_p, \sigma_q, \ell\right)\sigma_q^2}{2\sigma_p\sqrt{\sigma_p^2 + \sigma_q^2}}\right)\right),$$

y

$$\mathcal{I}_1\left(\mu_p, \sigma_p, \ell\right) = f\left(\mu_p, \sigma_p, \ell\right)\left(\sqrt{\pi} - \sqrt{\pi}\operatorname{erf}\left(\frac{d\left(\mu_p, \sigma_p, \ell\right)}{2\sqrt{2}}\right)\right), \qquad \text{(A.37)}$$

$$\mathcal{I}_1\left(\mu_q, \sigma_q, \ell\right) = f\left(\mu_q, \sigma_q, \ell\right)\left(\sqrt{\pi} - \sqrt{\pi}\operatorname{erf}\left(\frac{d\left(\mu_q, \sigma_q, \ell\right)}{2\sqrt{2}}\right)\right), \qquad \text{(A.38)}$$

$$\mathcal{I}_2\left(\mu_p, \sigma_p, \ell\right) = f_2\left(\mu_p, \sigma_p, \ell\right)\left(\sqrt{\pi} - \sqrt{\pi}\operatorname{erf}\left(\frac{d_2\left(\mu_p, \sigma_p, \ell\right)}{2\sqrt{2}}\right)\right), \qquad \text{(A.39)}$$

$$\mathcal{I}_2\left(\mu_q, \sigma_q, \ell\right) = f_2\left(\mu_q, \sigma_q, l\right)\left(\sqrt{\pi} - \sqrt{\pi}\operatorname{erf}\left(\frac{d_2\left(\mu_q, \sigma_q, \ell\right)}{2\sqrt{2}}\right)\right), \qquad \text{(A.40)}$$

$$d_1\left(\mu_p, \sigma_p, \ell\right) = \frac{\sqrt{2}\sigma_p}{\ell^2} = d_2\left(\mu_p, \sigma_p, \ell\right), \qquad \text{(A.41)}$$

$$d_1\left(\mu_q, \sigma_q, \ell\right) = \frac{\sqrt{2}\sigma_q}{\ell^2} = d_2\left(\mu_q, \sigma_q, \ell\right), \qquad \text{(A.42)}$$

$$f_1(\mu_p,\sigma_p,\ell) = \sigma_p^2\sqrt{\pi}\exp\left(-\frac{\left(2\ell^2\mu_p+\sigma_p^2\right)}{4\ell^4\sigma_p^2}\left(\sigma_p^2-2l^2\mu_p\right)-\frac{\mu_p^2}{\sigma_p^2}+\frac{d^2}{4}\right)$$
$$= f_2(\mu_p,\sigma_p,\ell), \tag{A.43}$$

$$f_1(\mu_q,\sigma_q,\ell) = \sigma_q^2\sqrt{\pi}\exp\left(-\frac{\left(2\ell^2\mu_q+\sigma_q^2\right)}{4\ell^4\sigma_q^2}\left(\sigma_q^2-2l^2\mu_q\right)-\frac{\mu_q^2}{\sigma_q^2}+\frac{d^2}{4}\right)$$
$$= f_2(\mu_q,\sigma_q,\ell), \tag{A.44}$$

$$\mathcal{I}(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell) = \frac{1}{2\pi\sigma_p\sigma_q}\left(\mathcal{I}_1(\mu_p,\sigma_q,\mu_p,\sigma_q,\ell)+\mathcal{I}_2(\mu_p,\sigma_q,\mu_p,\sigma_q,\ell)\right). \tag{A.45}$$

Similarmente

$$\mathcal{I}(\mu_p,\sigma_p,\ell) = \frac{1}{2\pi\sigma_p^2}\left(\mathcal{I}_1(\mu_p,\sigma_p,\ell)+\mathcal{I}_2(\mu_p,\sigma_p,\ell)\right), \tag{A.46}$$

y

$$\mathcal{I}(\mu_q,\sigma_q,\ell) = \frac{1}{2\pi\sigma_q^2}\left(\mathcal{I}_1(\mu_q,\sigma_q,\ell)+\mathcal{I}_2(\mu_q,\sigma_q,\ell)\right), \tag{A.47}$$

por lo tanto

$$\widehat{\gamma_k^2}(\mathbb{P},\mathbb{Q})_{Laplaciana} = \mathcal{I}(\mu_p,\sigma_p,\ell)+\mathcal{I}(\mu_q,\sigma_q,\ell)-2\mathcal{I}(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell)$$
$$= \frac{1}{2\pi\sigma_p^2}\left(\mathcal{I}_1(\mu_p,\sigma_p,\ell)+\mathcal{I}_2(\mu_p,\sigma_p,\ell)\right) \tag{A.48}$$
$$+ \frac{1}{2\pi\sigma_q^2}\left(\mathcal{I}_1(\mu_q,\sigma_q,\ell)+\mathcal{I}_2(\mu_q,\sigma_q,\ell)\right)$$
$$- \frac{1}{\pi\sigma_p\sigma_q}\left(\mathcal{I}_1(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell)+\mathcal{I}_2(\mu_p,\mu_q,\sigma_p,\sigma_q,\ell)\right). \ \square \tag{A.49}$$

Corolario 3.1.6 Si $k(\mathbf{x},\mathbf{y};\ell)$ es un kernel Laplaciano de dimensión n, es decir

$$k(\mathbf{x},\mathbf{y};\ell) = \exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|_1}{2\ell^2}\right), \quad \text{donde} \quad \mathbf{x},\ \mathbf{y}\in\mathbb{R}^n,$$

y

$$\widehat{p}(\mathbf{x}) = \frac{1}{|\mathbf{\Sigma}_p|^{1/2}(2\pi)^{n/2}}\exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu}_p)^\top\mathbf{\Sigma}_p^{-1}(\mathbf{x}-\boldsymbol{\mu}_p)}{2}\right) \quad \text{donde} \quad \mathbf{\Sigma}_p = diag\left(\sigma_{p1}^2,\sigma_{p2}^2,\cdots,\sigma_{pn}^2\right),$$
$$\widehat{q}(\mathbf{y}) = \frac{1}{|\mathbf{\Sigma}_q|^{1/2}(2\pi)^{n/2}}\exp\left(-\frac{(\mathbf{y}-\boldsymbol{\mu}_q)^\top\mathbf{\Sigma}_q^{-1}(\mathbf{y}-\boldsymbol{\mu}_q)}{2}\right) \quad \text{donde} \quad \mathbf{\Sigma}_q = diag\left(\sigma_{q1}^2,\sigma_{q2}^2,\cdots,\sigma_{qn}^2\right),$$

son estimadores de $p(x)$ y $q(y)$ respectivamente, donde los parámetros $\ell\in\mathbb{R}$, $\boldsymbol{\mu}_p = (\mu_{p1},\mu_{p2},\cdots,\mu_{pn})$, $\boldsymbol{\mu}_q = (\mu_{q1},\mu_{q2},\cdots,\mu_{qn})\in\mathbb{R}^n$ y $\mathbf{\Sigma}_p$, $\mathbf{\Sigma}_q\in\mathbb{R}^{n\times n}$ son matrices diagonales.

Entonces el estimador de la métrica $\gamma_k^2(\mathbb{P},\mathbb{Q})$ entre las distribuciones de probabilidad $\mathbb{P}$ y $\mathbb{Q}$ viene dado por la expresión

$$\begin{aligned}\widehat{\gamma_k^2}(\mathbb{P},\mathbb{Q}) &= \prod_{i=1}^{n}\frac{1}{2\pi\sigma_{pi}^2}\left(\mathcal{I}_1\left(\mu_{pi},\sigma_{pi},\ell\right)+\mathcal{I}_2\left(\mu_{pi},\sigma_{pi},\ell\right)\right)\\ &+ \prod_{i=1}^{n}\frac{1}{2\pi\sigma_{qi}^2}\left(\mathcal{I}_1\left(\mu_{qi},\sigma_{qi},\ell\right)+\mathcal{I}_2\left(\mu_{qi},\sigma_{qi},\ell\right)\right)\\ &- \prod_{i=1}^{n}\frac{1}{\pi\sigma_{pi}\sigma_{qi}}\left(\mathcal{I}_1\left(\mu_{pi},\mu_{qi},\sigma_{pi},\sigma_{qi},\ell\right)+\mathcal{I}_2\left(\mu_{pi},\mu_{qi},\sigma_{pi},\sigma_{qi},\ell\right)\right), \quad \text{(A.50)}\end{aligned}$$

donde las funciones $\mathcal{I}_1$ y $\mathcal{I}_2$ son definidas como en las ecuaciones 3.10 y 3.13 respectivamente.

Prueba Calculemos $\int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{x},\mathbf{y})p(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y}$.

$$\begin{aligned}\mathcal{I}\left(\boldsymbol{\mu}_p,\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_p,\boldsymbol{\Sigma}_p,\ell\right) = \int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{x},\mathbf{y})p(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} &= \int_{\mathcal{X}}\int_{\mathcal{X}}\left(\exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|_1}{2\ell^2}\right)\right)\\ &\times\left(\frac{1}{|\boldsymbol{\Sigma}_p|^{1/2}(2\pi)^{n/2}}\exp\left(-\frac{\left(\mathbf{x}-\boldsymbol{\mu}_p\right)^\top\boldsymbol{\Sigma}_p^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_p\right)}{2}\right)\right) \quad \text{(A.51)}\\ &\times\left(\frac{1}{|\boldsymbol{\Sigma}_q|^{1/2}(2\pi)^{n/2}}\exp\left(-\frac{\left(\mathbf{y}-\boldsymbol{\mu}_q\right)^\top\boldsymbol{\Sigma}_q^{-1}\left(\mathbf{y}-\boldsymbol{\mu}_q\right)}{2}\right)\right)d\mathbf{x}d\mathbf{y}.\end{aligned}$$

Redescribiendo la Ecuación A.51, obtenemos

$$\begin{aligned}\mathcal{I}\left(\boldsymbol{\mu}_p,\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_p,\boldsymbol{\Sigma}_p,\ell\right) = \int_{\mathcal{X}}\int_{\mathcal{X}}&\left(\prod_{i=1}^{n}\exp\left(\frac{-|x_i-y_i|}{2\ell^2}\right)\right)\\ &\times\left(\prod_{i=1}^{n}\frac{1}{\sigma_{pi}(2\pi)^{1/2}}\exp\left(-\frac{\left(x_i-\mu_{pi}\right)^2}{2\sigma_{pi}^2}\right)\right) \quad \text{(A.52)}\\ &\times\left(\prod_{i=1}^{n}\frac{1}{\sigma_{qi}(2\pi)^{1/2}}\exp\left(-\frac{\left(y_i-\mu_{qi}\right)^2}{2\sigma_{qi}^2}\right)\right)d\mathbf{x}d\mathbf{y}.\end{aligned}$$

Si suponemos $d\mathbf{x} = dx_1\cdot dx_2\cdots dx_n$ y $d\mathbf{y} = dy_1\cdot dy_2\cdots dy_n$ y $\mathcal{X} = \mathcal{X}_1\times\mathcal{X}_2\cdots\mathcal{X}_n$, entonces

$$\begin{aligned}\mathcal{I}\left(\boldsymbol{\mu}_p,\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_p,\boldsymbol{\Sigma}_p,\ell\right) = \prod_{i=1}^{n}\int_{\mathcal{X}_1\times\mathcal{X}_2\cdots\mathcal{X}_n}\int_{\mathcal{X}_1\times\mathcal{X}_2\cdots\mathcal{X}_n}&\left(\exp\left(\frac{-|x_i-y_i|}{2\ell^2}\right)\right)\\ &\times\left(\frac{1}{\sigma_{pi}(2\pi)^{1/2}}\exp\left(-\frac{\left(x_i-\mu_{pi}\right)^2}{2\sigma_{pi}^2}\right)\right) \quad \text{(A.53)}\\ &\times\left(\frac{1}{\sigma_{qi}(2\pi)^{1/2}}\exp\left(-\frac{\left(y_i-\mu_{qi}\right)^2}{2\sigma_{qi}^2}\right)\right)\left(dx_1\cdot dx_2\cdots dx_n\right)\left(dy_1\cdot dy_2\cdots dy_n\right).\end{aligned}$$

Organizando adecuadamente los terminos de la Ecuación $A.53$, obtenemos

$$\begin{aligned}\mathcal{I}\left(\boldsymbol{\mu}_p,\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_p,\boldsymbol{\Sigma}_p,\ell\right) = &\left(\int_{\mathcal{X}_1}\int_{\mathcal{X}_1}\left(\frac{1}{\sigma_{q1}(2\pi)^{1/2}}\exp\left(-\frac{\left(y_1-\mu_{q1}\right)^2}{2\sigma_{q2}^2}\right)\right)\right)\\ &\times\left(\int_{\mathcal{X}_2}\int_{\mathcal{X}_2}\left(\frac{1}{\sigma_{q2}(2\pi)^{1/2}}\exp\left(-\frac{\left(y_2-\mu_{q2}\right)^2}{2\sigma_{q2}^2}\right)\right)\right)\times\cdots\\ &\times\left(\int_{\mathcal{X}_n}\int_{\mathcal{X}_n}\left(\frac{1}{\sigma_{qn}(2\pi)^{1/2}}\exp\left(-\frac{\left(y_n-\mu_{qn}\right)^2}{2\sigma_{qn}^2}\right)\right)\right).\end{aligned} \tag{A.54}$$

Luego por el Terorema 3.1.5, para $i=1,2,\cdots,n$ se tiene que

$$\begin{aligned}\mathcal{I}\left(\mu_{pi},\mu_{qi},\sigma_{pi},\sigma_{pi},\ell\right) = &\int_{\mathcal{X}_i}\int_{\mathcal{X}_i}\left(\exp\left(\frac{-|x_i-y_i|}{2\ell^2}\right)\right)\\ &\times\left(\frac{1}{\sigma_{pi}(2\pi)^{1/2}}\exp\left(-\frac{\left(x_i-\mu_{pi}\right)^2}{2\sigma_{pi}^2}\right)\right)\\ &\times\left(\frac{1}{\sigma_{qi}(2\pi)^{1/2}}\exp\left(-\frac{\left(y_i-\mu_{qi}\right)^2}{2\sigma_{qi}^2}\right)\right)dx_i dy_i,\end{aligned} \tag{A.55}$$

por lo tanto

$$\mathcal{I}\left(\boldsymbol{\mu}_p,\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_p,\boldsymbol{\Sigma}_p,\ell\right) = \prod_{i=1}^{n}\mathcal{I}\left(\mu_{pi},\mu_{qi},\sigma_{pi},\sigma_{pi},\ell\right). \tag{A.56}$$

Por el Teorema 3.1.5, se tiene

$$\mathcal{I}\left(\mu_{pi},\mu_{qi},\sigma_{pi},\sigma_{pi},\ell\right) = \frac{1}{\pi\sigma_{pi}\sigma_{qi}}\left(\mathcal{I}_1\left(\mu_{pi},\mu_{qi},\sigma_{pi},\sigma_{qi},\ell\right)+\mathcal{I}_2\left(\mu_{pi},\mu_{qi},\sigma_{pi},\sigma_{qi},\ell\right)\right). \tag{A.57}$$

Por consigiente

$$\begin{aligned}\widehat{\gamma_k^2}\left(\mathbb{P},\mathbb{Q}\right) &= \prod_{i=1}^{n}\frac{1}{2\pi\sigma_{pi}^2}\left(\mathcal{I}_1\left(\mu_{pi},\sigma_{pi},\ell\right)+\mathcal{I}_2\left(\mu_{pi},\sigma_{pi},\ell\right)\right)\\ &+ \prod_{i=1}^{n}\frac{1}{2\pi\sigma_{qi}^2}\left(\mathcal{I}_1\left(\mu_{qi},\sigma_{qi},\ell\right)+\mathcal{I}_2\left(\mu_{qi},\sigma_{qi},\ell\right)\right)\\ &- \prod_{i=1}^{n}\frac{1}{\pi\sigma_{pi}\sigma_{qi}}\left(\mathcal{I}_1\left(\mu_{pi},\mu_{qi},\sigma_{pi},\sigma_{qi},\ell\right)+\mathcal{I}_2\left(\mu_{pi},\mu_{qi},\sigma_{pi},\sigma_{qi},\ell\right)\right).\end{aligned} \tag{A.58}$$

Appendix B

Pruebas de teoremas sobre el modelo autorregresivo basado en el método embebimiento de distribuciones de probabilidad en un RKHS

Este capítulo contiene las pruebas de los teoremas más importantes sobre el modelo autorregresivo basado en el método embebimiento de distribuciones de probabilidad en un RKHS. Probamos el Teorema 4.2.1 de la Sección 4.2 y el Teorema 4.3.1 de la Sección 4.3.

Teorema 4.2.1. Sea $\mathcal{C}_{X_iX_{i-k}}$ definido como en (4.8) y considere $\{(x_i^l, x_{i-j}^l)\}_{l=1}^{N_{xy}}$, para $j = 1, 2, \cdots, p$, diferentes conjuntos de muestras tomadas *i.i.d* de las distribuciones $\mathbb{P}(X_i, X_{i-j})$, entonces el estimador de $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_p]$ está dado por

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{A}\boldsymbol{\alpha} - \mathbf{b}\|_2^2. \tag{B.1}$$

Además, si $\left(\mathbf{A}^\top \mathbf{A}\right)^{-1}$ es invertible, entonces

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{A}\boldsymbol{\alpha} - \mathbf{b}\|_2^2 = \left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{b}, \tag{B.2}$$

donde $\mathbf{A} \in \mathbb{R}^{p \times p}$ con entradas $\{\mathrm{tr}(\widehat{\mathbf{K}}_i^\top \widehat{\mathbf{K}}_j)\}_{i=1,j=1}^{p,p}$, $\mathbf{b} \in \mathbb{R}^{p \times 1}$ con entradas $\{\mathrm{tr}(\mathbf{H}^\top \widehat{\mathbf{K}}_i)\}_{i=1}^{p}$, $\mathbf{H} \in \mathbb{R}^{N_{xy}p \times N_{xy}}$ es una matriz de bloques con bloques dados por $\{\mathbf{H}_k\}_{k=1}^{p}$ y $\widehat{\mathbf{K}}_i \in \mathbb{R}^{N_{xy}p \times N_{xy}}$ es una matriz de bloques con bloques dados por $\{\mathbf{K}_{k,i}\}_{k=1}^{p}$, con $\mathbf{H}_k = \mathbf{H}_{i-k,i} = \boldsymbol{\Upsilon}_{i-k}^\top \boldsymbol{\Phi}_i$ and $\mathbf{K}_{k,i} = \boldsymbol{\Upsilon}_k^\top \boldsymbol{\Upsilon}_i$.

Prueba. Sea $\{(x_i^l, x_{i-j}^l)\}_{l=1}^{N_{xy}}$, para $j = 1, 2, \cdots, p$, diferentes conjuntos de muestras tomadas *i.i.d* de las distribuciónes $\mathbb{P}(X_i, X_{i-j})$. Denotamos por $\boldsymbol{\Phi}_i$ la matriz de

característica construida a partir de los elementos $\{\phi(x_i^l)\}_{l=1}^{N_{xy}}$, y $\{\varphi(x_{i-j}^l)\}_{l=1}^{N_{xy}}$,

$$\mathbf{\Phi}_i = (\phi(x_i^1), \phi(x_i^2), \cdots, \phi(x_i^{N_{xy}})), \quad \mathbf{\Upsilon}_{i-j} = (\varphi(x_{i-j}^1), \varphi(x_{i-j}^2), \ldots, \varphi(x_{i-j}^{N_{xy}})).$$

Sean $\mathcal{C}_{X_i X_{i-k}}$ y $\mathcal{C}_{X_{i-j} X_{i-k}}$ los estimadores de los operadores de covarianza cruzada definidos como en [46]

$$\widehat{\mathbf{C}}_{X_i X_{i-k}} = \frac{1}{N_{xy}} \sum_{l=1}^{N_{xy}} \phi(x_i^l) \otimes \varphi(x_{i-k}^l) = \frac{1}{N_{xy}} \mathbf{\Phi}_i \mathbf{\Upsilon}_{i-k}^\top \tag{B.3}$$

$$\widehat{\mathbf{C}}_{X_{i-j} X_{i-k}} = \frac{1}{N_{xy}} \sum_{l=1}^{N_{xy}} \varphi(x_{i-j}^l) \otimes \varphi(x_{i-k}^l) = \frac{1}{N_{xy}} \mathbf{\Upsilon}_{i-j} \mathbf{\Upsilon}_{i-k}^\top. \tag{B.4}$$

Si reemplazamos las Ecuaciones B.3 y B.2 en la Ecuación (4.8), obtenemos

$$\mathbf{\Phi}_i \mathbf{\Upsilon}_{i-k}^\top = \sum_{j=1}^{p} \alpha_j \mathbf{\Upsilon}_{i-j} \mathbf{\Upsilon}_{i-k}^\top. \tag{B.5}$$

Ahora, si pre-multiplicamos la Ecuación (B.5) por $\mathbf{\Upsilon}_{i-k}^\top$, y pos-mumtiplicamos por $\mathbf{\Phi}_i$, entonces

$$\mathbf{\Upsilon}_{i-k}^\top \mathbf{\Phi}_i \mathbf{\Upsilon}_{i-k}^\top \mathbf{\Phi}_i = \sum_{j=1}^{p} \alpha_j \mathbf{\Upsilon}_{i-k}^\top \mathbf{\Upsilon}_{i-j} \mathbf{\Upsilon}_{i-k}^\top \mathbf{\Phi}_i. \tag{B.6}$$

Simplificando la Ecuación (B.6) obtenemos

$$\mathbf{\Upsilon}_{i-k}^\top \mathbf{\Phi}_i = \sum_{j=1}^{p} \alpha_j \mathbf{\Upsilon}_{i-k}^\top \mathbf{\Upsilon}_{i-j}. \tag{B.7}$$

Nosotros podemos escribir la Ecuación (B.7) como

$$\mathbf{H}_{i-k,i} = \sum_{j=1}^{p} \alpha_j \mathbf{K}_{i-k,i-j}, \tag{B.8}$$

donde $\mathbf{H}_{i-k,i} = \mathbf{\Upsilon}_{i-k}^\top \mathbf{\Phi}_i$, $\mathbf{K}_{i-k,i-j} = \mathbf{\Upsilon}_{i-k}^\top \mathbf{\Upsilon}_{i-j}$, y $k = 1, 2, \ldots, p$. Note que las entradas de la matriz $\mathbf{H}_{i-k,i}$ son productos puntos $\{\varphi(x_{i-k}^r)^\top \phi(x_i^s)\}_{r=1,s=1}^{N_{xy},N_{xy}}$. Estos productos puntos puden ser calculados usando una función kernel $\{h(x_{i-k}^r, x_i^s)\}_{r=1,s=1}^{N_{xy},N_{xy}}$. Del mismo modo, las entradas de $\mathbf{K}_{i-k,i-j}$ estan dadas por productos internos $\{\varphi(x_{i-k}^r)^\top \phi(x_{i-j}^s)\}_{r=1,s=1}^{N_{xy},N_{xy}}$, que también se puede calcular usando una función kernel $\{k(x_{i-k}^r, x_{i-j}^s)\}_{r=1,s=1}^{N_{xy},N_{xy}}$. Dado un conjunto de datos de series de tiempo y un valor para p, los valores de $\mathbf{H}_{i-k,i}$, y $\mathbf{K}_{i-k,i}$ dependen de los valores escogidos para i y N_{xy}.

Suponiendo que el proceso aleatorio de tiempo discreto es estacionario, podemos obtener una estimación para $\boldsymbol{\alpha}$ usando el siguiente conjunto de ecuaciones

$$\begin{bmatrix}\mathbf{H}_1\\ \mathbf{H}_2\\ \vdots\\ \mathbf{H}_p\end{bmatrix} = \begin{bmatrix}\mathbf{K}_{1,1} & \mathbf{K}_{1,2} & \cdots & \mathbf{K}_{1,p}\\ \mathbf{K}_{2,1} & \mathbf{K}_{2,2} & \cdots & \mathbf{K}_{2,p}\\ \vdots & \vdots & \vdots & \vdots\\ \mathbf{K}_{p,1} & \mathbf{K}_{p,2} & \cdots & \mathbf{K}_{p,p}\end{bmatrix}\begin{bmatrix}\alpha_1\mathbf{I}\\ \alpha_2\mathbf{I}\\ \vdots\\ \alpha_p\mathbf{I}\end{bmatrix}, \tag{B.9}$$

donde $\mathbf{I}$ es la matriz identidad de dimensión N_{xy}.

se puede encontrar un estimador para $\boldsymbol{\alpha}$ resolviendo la ecuación

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\|\mathbf{H}-\mathbf{K}\boldsymbol{\alpha}_{N_{xy}}\right\|_2^2 = \operatorname{tr}\left((\mathbf{H}-\mathbf{K}\boldsymbol{\alpha}_{N_{xy}})^\top(\mathbf{H}-\mathbf{K}\boldsymbol{\alpha}_{N_{xy}})\right), \tag{B.10}$$

donde $\mathbf{H}\in\mathbb{R}^{p\times N_{xy}}$ es una matriz de bloques, donde los bloques están definidos por $\{\mathbf{H}_k\}_{k=1}^p$; $\mathbf{K}\in\mathbb{R}^{N_{xy}p\times N_{xy}p}$ es una matriz de bloques donde los bloques están definidos por $\{\mathbf{K}_{k,j}\}_{k=1,j=1}^{p,p}$; y $\boldsymbol{\alpha}_{N_{xy}}\in\mathbb{R}^{N_{xy}p\times N_{xy}}$ es también una matriz de bloques donde los bloques están definidos por $\{\alpha_k\mathbf{I}\}_{k=1}^p$. Por conveniencia, también definimos $\widehat{\mathbf{K}}_i\in\mathbb{R}^{N_{xy}p\times N_{xy}}$ como una matriz de bloque tomada de $\mathbf{K}$, con bloques dados por $\{\mathbf{K}_{k,i}\}_{k=1}^p$.

Finalmente, el problema de optimizar (B.10) es equivalente a el problema de mínimos cuadrados

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{A}\boldsymbol{\alpha}-\mathbf{b}\|_2^2, \tag{B.11}$$

donde $\mathbf{A}\in\mathbb{R}^{p\times p}$ con entradas $\{\operatorname{tr}(\widehat{\mathbf{K}}_i^\top\widehat{\mathbf{K}}_j)\}_{i=1,j=1}^{p,p}$, y $\mathbf{b}\in\mathbb{R}^{p\times 1}$ con entradas $\{\operatorname{tr}(\mathbf{H}^\top\widehat{\mathbf{K}}_i)\}_{i=1}^p$. □

Teorema 4.3.1. Si x_i^* minimiza la expresión

$$x_i^* = \arg\min_x f(x) = \arg\min_x \left\|\sum_{j=1}^p \alpha_j\varphi(x_{i-j})\otimes\phi(x)-\varphi(x)\otimes\phi(x)\right\|^2_{\mathcal{H}_1\otimes\mathcal{H}_2},$$

entonces

$$x_i^* = \frac{\sum_{j=1}^p \alpha_j k(x_{i-j},x_i^*)x_{i-j}}{\sum_{k=1}^p \alpha_k k(x_{i-k},x_i^*)}. \tag{B.12}$$

Prueba. Si aplicamos producto tensor a ambos lados de la expresión (4.19), obtenemos

$$\tau_i^*\otimes\phi(x_i^*) = \sum_{j=1}^p \alpha_j\varphi(x_{i-j})\otimes\phi(x_i^*). \tag{B.13}$$

Para encontrar un estimador de x_i^*, nosotros minimizamos en el espacio $\mathcal{H}_1\otimes\mathcal{H}_2$ la expresión

$$x_i^* = \arg\min_x f(x) = \arg\min_x \left\|\sum_{j=1}^p \alpha_j\varphi(x_{i-j})\otimes\phi(x)-\varphi(x)\otimes\phi(x)\right\|^2_{\mathcal{H}_1\otimes\mathcal{H}_2},$$

donde hemos definido

$$f(x) = \left\| \sum_{j=1}^{p} \alpha_j \varphi(x_{i-j}) \otimes \phi(x) - \varphi(x) \otimes \phi(x) \right\|^2_{\mathcal{H}_1 \otimes \mathcal{H}_2}. \tag{B.14}$$

La expresión $f(x)$ puede también ser escrita como

$$\begin{aligned} f(x) = & \left\langle \sum_{j=1}^{p} \alpha_j \varphi(x_{i-j}) \otimes \phi(x), \sum_{k=1}^{p} \alpha_k \varphi(x_{i-k}) \otimes \phi(x) \right\rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2} \\ & - 2 \left\langle \sum_{j=1}^{p} \alpha_j \varphi(x_{i-j}) \otimes \phi(x), \varphi(x) \otimes \phi(x) \right\rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2} \\ & + \langle \varphi(x) \otimes \phi(x), \varphi(x) \otimes \phi(x) \rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2}. \end{aligned} \tag{B.15}$$

Utilizando la propiedad $\langle u \otimes v, a \otimes b \rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2} = \langle u \otimes a \rangle_{\mathcal{H}_1} \langle v \otimes b \rangle_{\mathcal{H}_2}$, obtenemos

$$\begin{aligned} f(x) = & \left\langle \sum_{j=1}^{p} \alpha_j \varphi(x_{i-j}), \sum_{k=1}^{p} \alpha_k \varphi(x_{i-k}) \right\rangle_{\mathcal{H}_1} \langle \phi(x), \phi(x) \rangle_{\mathcal{H}_2} \\ & - 2 \left\langle \sum_{j=1}^{p} \alpha_j \varphi(x_{i-j}), \varphi(x) \right\rangle_{\mathcal{H}_1} \langle \phi(x), \phi(x) \rangle_{\mathcal{H}_2} \\ & + \langle \varphi(x), \varphi(x) \rangle_{\mathcal{H}_1} \langle \phi(x), \phi(x) \rangle_{\mathcal{H}_2}. \end{aligned} \tag{B.16}$$

Note que $C = \left\langle \sum_{j=1}^{p} \alpha_j \varphi(x_{i-j}), \sum_{k=1}^{p} \alpha_k \varphi(x_{i-k}) \right\rangle_{\mathcal{H}_1}$ es una constante (no depende de x), y los kernels $k(x, x')$ son de la forma $g(\|x - x'\|^2)$. Simplificando la expresión (B.16), obtenemos

$$f(x) = Cg(0) - 2g(0) \sum_{j=1}^{p} \alpha_j k(x_{i-j}, x) + g^2(0). \tag{B.17}$$

Tomando derivada con respecto a x, se obtiene la ecuación

$$\frac{df(x)}{dx} = -2g(0) \sum_{j=1}^{p} \alpha_j \frac{dk(x_{i-j}, x)}{dx}. \tag{B.18}$$

Si usamos la función kenel base radial (RBF)

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\ell \|\mathbf{x} - \mathbf{x}'\|_2^2\right), \tag{B.19}$$

donde ℓ^2 se conoce como el ancho de banda, de la expresión (B.18) se obtiene

$$\frac{df(x)}{dx} = -\frac{2g(0)}{\ell^2} \sum_{j=1}^{p} \alpha_j k(x_{i-j}, x)(x_{i-j} - x). \tag{B.20}$$

Igualando a cero y resolviendo para x, obtenemos la siguiente ecuación de punto fijo

$$x_i^* = \frac{\sum_{j=1}^{p} \alpha_j k(x_{i-j}, x_i^*) x_{i-j}}{\sum_{k=1}^{p} \alpha_k k(x_{i-k}, x_i^*)}. \quad \square \tag{B.21}$$